

University of Oxford
Department of Statistics Summer Project



MCMC and the Infinite Sites Model

Miklós Zoltán Rácz

Advisors:

Jotun Hein
Rune Lyngsø
István Miklós

University of Oxford
Department of Statistics

August 23, 2009

Contents

1	Introduction	1
2	The Infinite Sites Model	2
2.1	Basic terminology	2
2.2	Assumptions of the infinite sites model	3
2.3	Probability of evolutionary histories	4
3	MCMC on the Infinite Sites Model	6
3.1	Markov chain on the evolutionary histories by swapping consecutive events	7
3.2	Metropolis-Hastings algorithm	11
4	Mixing properties	11
4.1	Techniques to prove rapid mixing	13
4.2	A similar problem: generating linear extensions of a poset	14
4.3	Example when path coupling works	15
4.4	Lower bound on the mixing rate	19
5	Conclusions and Future Work	19
A	Acceptance Ratio Can Be Small	20
	References	22

1 Introduction

Coalescent theory, originally developed by Kingman [23], is a much researched retrospective model in population genetics that seeks to understand aspects of the evolutionary past of a sample of DNA sequences through analysis of the present day sample. An important quantity to calculate in order to infer biologically relevant quantities—such as the mutation rate—is the probability that they evolved from a known ancestor.

The infinite sites model, which describes the evolution of infinitely long sequences, is commonly used to model mutations in sequences. In a series of papers Ethier, Griffiths, and Tavaré [12, 14] published a recursion for calculating these probabilities under the infinite sites model. However, implementing this recursion for large data sets is slow, and therefore approximations are needed in order to accelerate calculations while still being able to make estimates of the biologically relevant parameters. Previous approximation

methods include corner-cutting [19] and stochastic methods such as importance sampling [4, 18, 29]. Other stochastic methods such as Markov chain Monte Carlo have been used to study population histories and evolutionary processes [13, 24, 30, 31], however these studies used different (and more difficult, complex) models than the infinite sites model. We are not aware of work studying Markov chain Monte Carlo sampling under the infinite sites model.

In this report—a summary of a six-week long undergraduate research project at the Department of Statistics, University of Oxford—we look at Markov chain Monte Carlo methods that sample from the evolutionary histories of a data set under the infinite sites model assumption. In the following section we introduce the basic terminology and the infinite sites model. In Section 3 we introduce the irreducible Markov chain that we study, and in Section 4 we discuss techniques to prove rapid mixing of this Markov chain and relate our problem to the problem of generating linear extensions of a partially ordered set. Finally, we summarize our results and look forward to what needs to be done in the future.

2 The Infinite Sites Model

2.1 Basic terminology

In the following a sequence is assumed to be a string or section of DNA. A particular sequence may undergo a duplication event, in which an exact copy of itself is created, resulting in two identical sequences. Alternatively, at certain points along the sequence, mutations may occur due to various causes (such as copying errors in the genetic material during cell division or exposure to ultraviolet light). These mutations alter the genetic information at that point in the sequence. Positions where two sequences differ due to a mutation are called segregating sites. We often view events backwards in time; in this case we refer to a duplication event as a coalescence event and to a mutation as a back mutation. Two sequences can also take part in a recombination event; however, in this model we are concerned only with duplication and mutation events, and we therefore assume the absence of recombination (see also Section 2.2).

The most recent common ancestor (MRCA) of a sample of sequences is the first sequence encountered when viewed backwards in time, which is an ancestor of all sequences present in the sample. A series of events which leads from the MRCA to the sample is called an evolutionary history (EH). Usually there are many different possible orders of duplication and mutation

events that generate the observed sample from the MRCA; these make up the set of possible EHs. Mutations are measured relative to the MRCA, i.e. we assume that no mutations are present in the MRCA. Consequently, the earliest event to occur in the history of the observed sample will be a duplication event. In this report we assume that the MRCA is known. A possible intermediate stage between the MRCA and the observed sample is called an ancestral configuration (AC). Two ACs are said to be in the same generation if they must undergo the same number of events backwards in time (coalescences or back mutations) in order to reach their MRCA. See Figure 1 for an illustration.

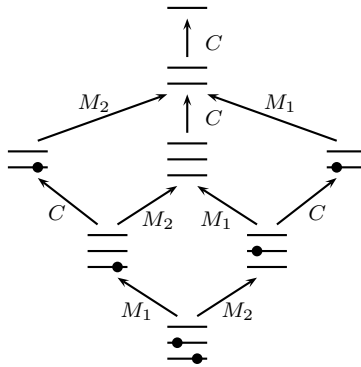


Figure 1: The configuration network, representing the possible evolutionary histories of a sample of 3 sequences with 2 segregating sites. The sample is shown at the bottom of the figure, while the MRCA is the top-most sequence. Mutations are represented by black dots. Here evolutionary histories are viewed backwards in time, hence the labels C (coalescence) and M (back mutation).

For more on coalescent theory see [17].

2.2 Assumptions of the infinite sites model

The infinite sites model (ISM), first proposed by Kimura in 1969 [22], models the evolution of very long (practically infinite) sequences. Each sequence consists of infinitely many sites, each of which mutates with the same, infinitesimally low probability; the rate of a mutation occurring anywhere among the sequences is denoted by θ and called the scaled mutation rate. Since this rate is finite, only finitely many sites in the sample are segregating sites. Furthermore, due to the infinite length assumption, whenever a mutant appears,

it represents a mutation at a new site, thus increasing the number of segregating sites; the probability of a site mutating more than once is negligible. Therefore at every site only two states are possible: ancestral or mutated, and consequently every sequence can be represented by a finite binary string. Each digit in the string corresponds to a segregating site; 0 denotes an ancestral state while 1 denotes a mutated state. The model assumes the absence of recombination.

2.2.1 Representation of a data set

We assume that the sequences are unlabelled—i.e. we do not distinguish between identical sequences, we only know how many of a particular type there is—and that the mutations are labelled by position. In this case we can represent the data by a matrix S and a vector \mathbf{v} . If there are k different types of sequences and m segregating sites, then S is a matrix with k rows representing the different types of sequences and m columns representing the positions of the segregating sites. An element s_{ij} of the matrix is 1 if sequence type i carries a mutation in position j , and 0 if it is in the ancestral state (where $i = 1, \dots, k$ and $j = 1, \dots, m$). The vector $\mathbf{v} = (n_1, n_2, \dots, n_k)$ gives the multiplicities of each sequence type. We denote the total number of sequences by $n = \sum_{i=1}^k n_i$.

2.3 Probability of evolutionary histories

We are interested in the following: we are given a data set $D = (S, \mathbf{v})$ which consists of a sample of sequences. These samples share a genealogical history, i.e. a line of descent from their MRCA, which we assume is known. However, there are many possible EHs, each with different probabilities; some are more likely than others. In order to infer biologically relevant quantities, such as the mutation rate, it is essential to be able to integrate over the possible EHs according to their probabilities.

In a series of papers from 1987 to 1995, Ethier, Griffiths, and Tavaré proposed a recursion formula for calculating the probability $\mathbb{P}_\theta(S, \mathbf{v})$ of a set of sequences (S, \mathbf{v}) given that it was generated under the infinite sites model. The formula can be obtained by a backward-forward argument: we look at what could have been the last event in the evolutionary history (either a duplication or a mutation event) and decompose the probability according to this; then consider what is the probability of reaching the current sample

from a possible previous configuration. The recursion takes the form

$$\begin{aligned} \mathbb{P}_\theta(S, \mathbf{v}) &= \frac{n-1}{n-1+\theta} \sum_{i:n_i>1} \frac{n_i-1}{n-1} \mathbb{P}_\theta(S, \mathbf{v} - \mathbf{e}_i) \\ &+ \frac{\theta}{n-1+\theta} \sum_{i:\text{singleton}} \frac{n_{i'}+1-\delta_{i,i'}}{n} \mathbb{P}_\theta(\tilde{S}, \mathbf{v} - \mathbf{e}_i + \mathbf{e}_{i'}) \end{aligned}$$

with initial condition $\mathbb{P}_\theta([0, \dots, 0], \{1\}) = 1$. In the above formula $n = \sum_{i=1}^k n_i$ is the number of sequences in the sample, \mathbf{e}_i is a unit vector of length k whose i^{th} entry is 1. Singleton sites are those segregating sites where only one sequence in the sample carries the mutated state at that position and all the other sequences carry the ancestral state. Singletons are those sequences which carry a mutation at a singleton site. In the formula above we denote the index of the sequence obtained when this singleton mutation is removed by i' . If the sequence produced after the mutation is removed is unique within the sample then we let $i' = i$. We denote the new matrix obtained after a back mutation by \tilde{S} . For a more detailed explanation of this recursion, see Chapter 2.4.2 of [17]. Unfortunately this recursion is slow for large data sets and therefore other methods must be used in these cases.

The probability of a data set $D = (S, \mathbf{v})$ can also be calculated as

$$\mathbb{P}_\theta(D) = \sum_{\mathcal{H}} \mathbb{P}_\theta(\mathcal{H}), \quad (2.1)$$

where the summation is over all possible EHs \mathcal{H} that lead from the MRCA of the sequences in D to D . The probability of a particular EH \mathcal{H} can be calculated by going backwards in time from the data set D to the MRCA and multiplying the appropriate factors appearing in the EGT recursion, depending on what the next event (coalescence or back mutation) is. Suppose we currently have l sequences ($2 \leq l \leq n$, where n is the number of sequences in the data set). The appropriate factors are then the following in the different cases:

- If the event is the coalescence of two sequences of type i and the multiplicity of type i is l_i before the coalescent event, then the factor is:

$$\frac{l-1}{l-1+\theta} \frac{l_i-1}{l-1} = \frac{l_i-1}{l-1+\theta}.$$

- If the event is a back mutation, then there are two cases to consider. Either the new type after the back mutation can be different of all other types present in the sample, in which case the factor is:

$$\frac{\theta}{(l-1+\theta)l};$$

or the new type after the back mutation can be identical to another type in the sequence. If this other type is type i with multiplicity l_i before the back mutation, then the factor is:

$$\frac{\theta(l_i + 1)}{(l - 1 + \theta)l}.$$

Thus for a particular EH \mathcal{H} calculating $\mathbb{P}_\theta(\mathcal{H})$ is easy—the problem with calculating $\mathbb{P}_\theta(D)$ according to 2.1 is that the number of possible EHs can grow exponentially in the input size, which in our case is the number of events in an EH, which is $n + m - 1$.

For these reasons we approach the problem via stochastic methods; specifically, we use MCMC methods: we construct an irreducible Markov chain on the possible EHs of a data set.

3 MCMC on the Infinite Sites Model

As mentioned previously, it is important to be able to sample from the evolutionary histories of a given sample of sequences according to their probabilities in order to estimate relevant biological parameters. Previous work mainly focused on importance sampling methods. Importance sampling [25] requires a proposal distribution to be defined over the histories that is easy to sample from. In order for importance sampling to be effective, the proposal distribution has to be similar to the actual distribution we want to sample from. Several proposal distributions have been proposed in the literature, see [4, 18, 29].

Another widely used class of Monte Carlo sampling techniques are Markov Chain Monte Carlo (MCMC) methods [25]. MCMC methods have been used to study population histories and evolutionary processes [13, 24, 30, 31], however these studies used different models than the infinite sites model. We are not aware of work studying MCMC sampling under the ISM. In the following we develop a MCMC method to sample from the evolutionary histories of a given sample of sequences under the infinite sites model.

In the following we look at the events backwards in time, so an EH will be a sequence of coalescent and back mutation events starting with the data set and ending with the MRCA. In this section we introduce the Markov chain on the EHs that we study and we show that it is irreducible on the set of possible EHs of a set of sequences. In order to be able to sample from the very large number of possible EHs, it is necessary for the Markov chain to mix rapidly. This will be discussed in Section 4.

3.1 Markov chain on the evolutionary histories by swapping consecutive events

We define an irreducible and aperiodic Markov chain on the EHs that has the uniform distribution as its stationary distribution. In order to ensure aperiodicity, at every step of the Markov chain we flip a coin; with probability $1/2$ we stay in the current state and with probability $1/2$ we do the following: suppose the Markov chain is at a particular state x (i.e. a particular EH). Pick one of the ACs of x uniformly at random, excluding the MRCA and the data set itself. If there are n sequences in the sample, and m segregating sites altogether, then there are altogether $n + m - 1$ events during an EH of the sample, and thus there are $n + m - 2$ ACs if we exclude the MRCA (the final AC, if viewed backwards in time) and the data set (the first AC, if viewed backwards in time), so pick any AC with probability $1/(n + m - 2)$. Now look at the two consecutive events of x that are “around” this chosen AC, i.e. the event that is right before and the one that is right after the chosen AC. These events are swappable if swapping the two events (i.e. changing the order of these two events and leaving all other events in x unchanged) is possible, that is it leads to another possible EH of the data set. If the two events are swappable, then swap them, otherwise leave the EH unchanged. See Figure 2 for an illustration.

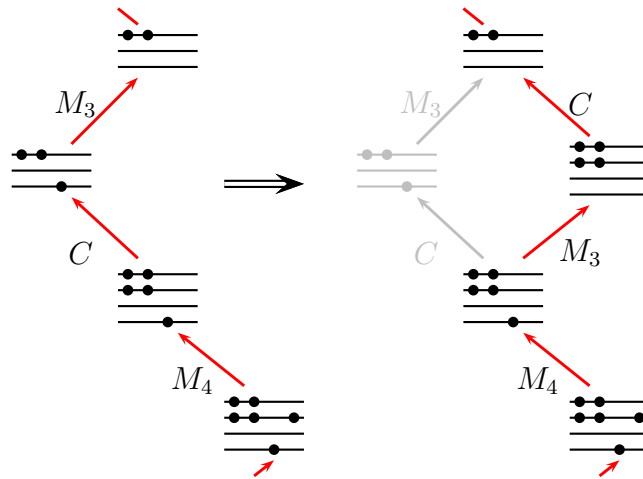


Figure 2: Swapping two consecutive events of an evolutionary history.

Let us call two EHs neighbors if a swap described above can take one into the other. Denote the number of neighbors of an EH x by $N(x)$. The

transition matrix of this Markov chain then has entries

$$P_{xy} = \begin{cases} \frac{1}{2(n+m-2)} & \text{if } x \text{ and } y \text{ are neighbors,} \\ 1 - \frac{N(x)}{2(n+m-2)} & \text{if } y=x, \\ 0 & \text{otherwise.} \end{cases}$$

One can see that this chain is reversible with respect to the uniform distribution, and therefore its stationary distribution is the uniform distribution.

3.1.1 Irreducibility

To show that the above Markov chain is irreducible, let us consider two different EHs of the data set, x and y . Our task is to show that if the Markov chain is in x then after a number of steps it is possible for the Markov chain to be in y . A possible step, as defined above, is swapping consecutive events of an EH.

The ACs are divided into generations: two ACs are in the same generation if they must undergo the same number of events in order to reach their MRCA. We number these generations going backwards in time, for convenience: an AC is in generation g if it has to undergo $n + m - 1 - g$ events to reach the MRCA. In particular, the data set is in generation 0, and the MRCA is in generation $n + m - 1$. Two evolutionary histories agree until generation g if their ACs are the same until generation g but their ACs in generation $g + 1$ are different. Suppose the evolutionary histories x and y agree until generation g (where $0 \leq g \leq n + m - 4$ in order for them to be different), and let us call this ancestral configuration AC_g . By induction on g , it is enough to prove that just by swapping consecutive events we can make the evolutionary histories x and y agree until generation $g + 1$.

The basic idea is the following (illustrated in Figure 3). Suppose the event after generation g in y is a back mutation at site j for $1 \leq j \leq m$ (denoted by M_3 in the figure). In this case there is exactly one type of sequence with a 1 in site j among the sequences found in AC_g , and this type has multiplicity 1 in AC_g . In order for x to reach the MRCA, there has to be a back mutation at site j somewhere along the events of x . This event is swappable with all previous events one after another along x up until generation g , since none of the previous events have to have happened before this back mutation event could take place. The reason for this is that in order for this back mutation at site j to happen, there must be exactly one type of sequence with a 1 in site j , and this type has to have multiplicity 1; we know that this is true in AC_g , and coalescences and back mutation events at other sites will not change this fact. Therefore swapping consecutive events along x we can “bring down” the

back mutation event at site j on x and make x and y agree until generation $g + 1$.

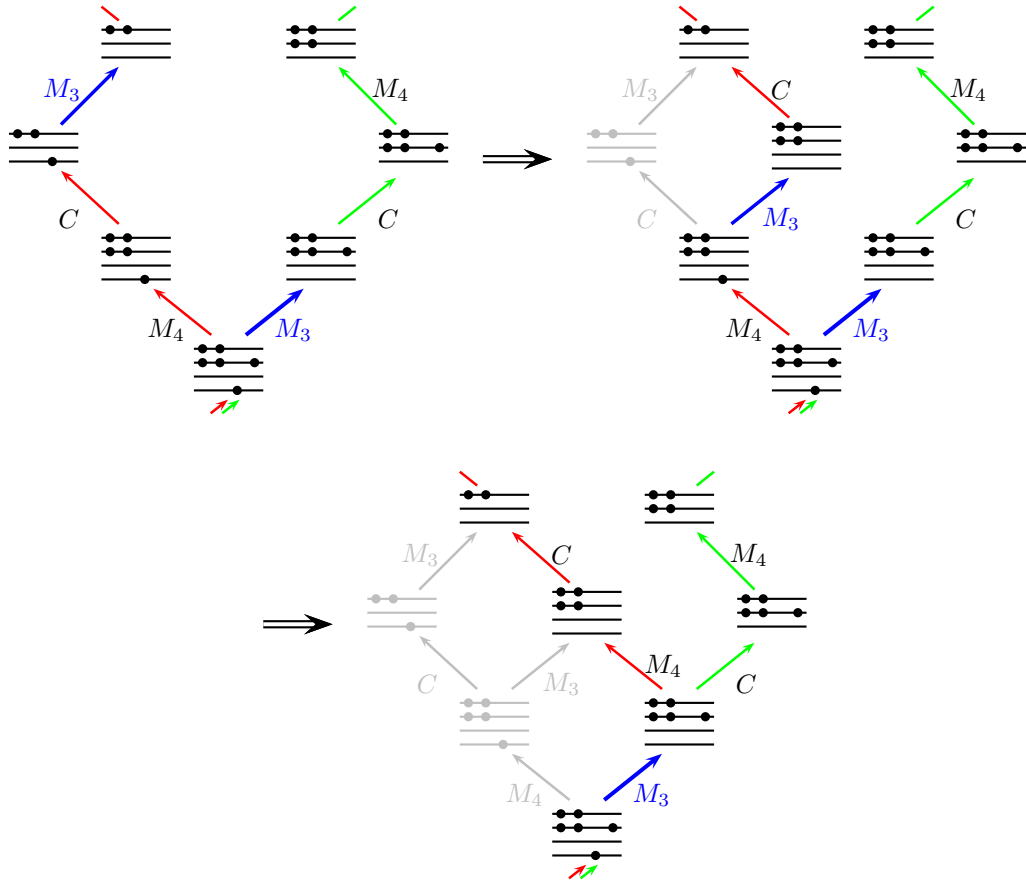


Figure 3: A path by swapping consecutive events that makes the two evolutionary histories agree in an additional AC.

If the event after generation g in y is a coalescence of two sequences of type i for $1 \leq i \leq k$, then the same idea works. In this case there are multiple sequences of type i in AC_g . In order for x to reach the MRCA, there has to be a coalescence of two sequences of type i somewhere along x after generation g . Consider the first event of this type along x (backwards in time) and call this C_i . This event is swappable with all previous events one after another along x up until generation g , since none of the previous events have to have happened before C_i could take place. The reason for this is that in order for C_i to happen, there must be multiple sequences of type i ; we know that there were multiple sequences of this type in AC_g , and until a coalescent event of this type happens, there will always be multiple sequences of this

type. Therefore swapping consecutive events along x we can “bring down” the coalescent event C_i on x and make x and y agree until generation $g + 1$.

Thus in this way by swapping consecutive events we can make x and y agree until generation $g + 1$, and therefore we are done by induction.

The constructed path is also (one of) the shortest path(s) between x and y if only swaps of consecutive events are allowed. First of all, a lower bound for the shortest path between x and y is the number of pairs of events that are in one order in x and another order in y , since a swap of consecutive events can decrease this number by at most one. On the other hand, it follows from the above construction that during every swap we swap a pair of events that are not in the same order in x and y , and thus we decrease this number by one in every step.

3.1.2 Variants

Variants of the Markov chain above can also be considered. For instance, instead of picking two consecutive events at random, choose two events of the EH at random, and if these are swappable then swap them, otherwise leave the EH unchanged. Since any move possible in the previous Markov chain is possible here as well, this Markov chain is irreducible too.

Another possible variant can be constructed based upon the above construction of a path between two EHs. Choose two events of the current EH x at random and consider all the events in between. These events can be rotated if the following modification of x gives a possible EH of data set: move the oldest event to the place of the most recent event, and “move up” all other events a generation closer to the MRCA. If the events can be rotated, then rotate them, otherwise leave x unchanged. If we choose two consecutive events, then swapping and rotating them is the same thing, and therefore any move that is possible in the original Markov chain is possible here as well, and consequently this Markov chain is irreducible too. The idea for this Markov chain comes from the proof of the irreducibility of the original Markov chain – a move of rotating events is basically just multiple moves of swapping consecutive events put together.

We focus on the original Markov chain, however, because we suspect that these variants have roughly the same mixing properties. We do however introduce a modification in Section 4: in choosing an AC of the current EH, we do not have to restrict ourselves to choosing according to the uniform distribution, we may choose an arbitrary distribution. A specific distribution will be chosen in Section 4.

3.2 Metropolis-Hastings algorithm

The constructed Markov chain converges to the uniform distribution on the possible EHs of the data set; however, the biologically relevant distribution is the one that is described in Section 2.3. Using the Metropolis-Hastings algorithm we can create a modified Markov chain whose stationary distribution is the desired distribution. Metropolis et al. [26] introduced the fundamental idea of modifying a Markov chain in order to be able to generate random samples from a desired distribution π , which Hastings later generalized [16]. The algorithm consists of two steps. Suppose the Markov chain is in state x at some time t . First, generate a state y according to the original transition matrix P . Second, accept the move to state y with probability $\alpha_{xy} = \min\{1, \frac{\pi(y)P_{yx}}{\pi(x)P_{xy}}\}$ (called the acceptance probability), otherwise stay in state x . In our case P is symmetric and therefore the acceptance probabilities are $\alpha_{xy} = \min\{1, \frac{\pi(y)}{\pi(x)}\}$.

In this way we have constructed an irreducible Markov chain on the EHs that converges to the desired distribution. While it remains to be shown that this Markov chain mixes rapidly, we now attempt to prove that the originally presented Markov chain mixes rapidly. The modified Markov chain presents difficulties because the acceptance probabilities can be very small, as small as $(c_1 n)^{-c_2 m}$, where c_1 and c_2 are positive constants, possibly causing bottlenecks in the chain (see Appendix A for detailed calculations). While dealing with the chain with uniform stationary distribution may not have biological relevance in itself, it serves as a stepping stone to the more difficult, biologically relevant problem. Therefore, we consider this chain in the next section.

4 Mixing properties

In analyzing the efficiency of random sampling of combinatorial structures from a specified distribution via Markov chain simulation, the key issue is to determine the *mixing rate* of the chain, i.e. the number of simulation steps needed to ensure that it is sufficiently close to its equilibrium distribution. In order for the algorithm to be efficient this number has to be reasonably small, which usually means drastically less than the size of state space itself. In many combinatorial problems—such as the one we are concerned with now—the size of the state space of the Markov chain typically grows exponentially with the input size. However, in order for the algorithms to be efficient we require the mixing rate to be bounded by a polynomial in the input size. Informally, we say such Markov chains are *rapidly mixing*. If there is an

exponential lower bound for the mixing rate then we say that the chain is *torpid mixing*.

In the following, we denote the state space of the Markov chain by Ω and its size by $N = |\Omega|$. The transition matrix of the chain on Ω is denoted by P , and we assume that P is irreducible and reversible with respect to the probability distribution π on Ω , i.e. the detailed balance condition holds:

$$\pi(x) P_{xy} = \pi(y) P_{yx}.$$

This implies that π is the stationary distribution of P . If in addition P is aperiodic, then the distribution of the state at time t converges pointwise to π as $t \rightarrow \infty$, regardless of initial state. In this case the chain is called ergodic.

It is well known from Perron-Frobenius theory that P has real eigenvalues $1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_N \geq -1$. If P is ergodic then $\lambda_N > -1$, and the rate of convergence to π is governed by the second-largest eigenvalue in absolute value, $\lambda_{max} = \max\{\lambda_2, |\lambda_N|\}$. The value $1 - \lambda_{max}$ is called the spectral gap. In order to not have to pay attention to λ_N , it is common to define a lazy walk, where we flip a coin at every step of the Markov chain, and with probability $1/2$ we stay at the current state and with probability $1/2$ we move according to the chain. By this method, all the eigenvalues become non-negative and we only have to deal with λ_2 , while the spectral gap decreases only by a factor of 2.

There are various ways to quantify the rate of convergence of a Markov chain to its equilibrium distribution, here we only look at a few examples. One common way is to look at the *relaxation time* defined by

$$\tau_{rel} = \frac{1}{1 - \lambda_2}.$$

Another measure is defined by Sinclair in [27]. First, we define the *total variation distance* between probability measures. Suppose μ and ν are two probability measures on Ω . Their total variation distance is defined as

$$d_{TV}(\mu, \nu) = \max_{S \subseteq \Omega} |\mu(S) - \nu(S)| = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

If the Markov chain is in state x at time $t = 0$ then the distribution of the state at time t is $\delta_x^T P^t$, where δ_x is the Dirac delta which puts all its mass on x . Sinclair measures the rate of convergence using the function τ_x which is defined for $\varepsilon > 0$ by

$$\tau_x(\varepsilon) = \min\{t : d_{TV}(\delta_x^T P^t, \pi) \leq \varepsilon \text{ for all } t' \geq t\}.$$

For a good description of measures quantifying the rate of convergence, including inequalities that describe the relationships between these parameters, see Aldous and Fill’s book [3].

4.1 Techniques to prove rapid mixing

Various techniques exist in the literature to prove the rapid mixing of Markov chains, such as geometric arguments, coupling, and group representation theory. Here we do not go into detail about these methods, but merely provide pointers to literature on the subject. We will describe the specific technique of path coupling in detail in Section 4.3 when proving the rapid mixing of the considered Markov chain in a special case.

Geometric arguments are mainly concerned with formalizing the intuition that if there is a bottleneck separating two parts of the state space, then this prevents rapid convergence to the steady state; however, if no such bottlenecks are present, then the Markov chain mixes rapidly. The techniques focus on bounding the second largest eigenvalue λ_2 , using conductance, canonical path, and multicommodity flow arguments. For more details and precise statements, see Jerrum and Sinclair [20, 28], Diaconis and Stroock [9], and Sinclair [27].

Coupling is a technique that generated from Doeblin [10] from 1938, which was revived in Markov chain theory in the 1970s. The main starting point is the “Coupling Lemma” (see e.g. [1]):

Lemma 4.1. *Let (X, Y) be a random process (the coupling) such that marginally both X and Y are copies of the same Markov chain M . Moreover, suppose that Y_0 is chosen from the stationary distribution π of M , and μ_t is the distribution of X_t . Then*

$$d_{TV}(\mu_t, \pi) \leq \mathbb{P}(X_t \neq Y_t).$$

When $X_t = Y_t$, we say that X and Y have *coupled*. Thus we can see from the Coupling Lemma that the goal is to create a joint process on two copies of the Markov chain, a coupling, that will have a probabilistic tendency to come together quickly. There have been various approaches to the coupling technique, such as the *path coupling* technique developed by Bubley and Dyer [6]. A more detailed description of the path coupling technique will follow in Section 4.3 when we use it to prove the rapid mixing of the considered Markov chain in a special case.

Group representation theory is a powerful tool when analyzing random walks on structures with a high degree of symmetry, such as card shuffling [2]. See the book by Diaconis [8] for an excellent survey.

4.2 A similar problem: generating linear extensions of a poset

The problem of generating an EH of a sample of sequences uniformly at random is very similar to the problem of generating linear extensions of a partially ordered set (poset) uniformly at random.

In this section we deviate from our original notations and P , n , and N will denote new quantities. We return to our original notations in the next section.

Let $N = \{1, 2, \dots, n\}$ and $P = (N, \preceq)$ be a partial order. A linear extension of P is a total order $X = (N, \sqsubseteq)$ which respects P , i.e. $i \preceq j$ implies $i \sqsubseteq j$ for all $i, j \in N$. Let $\Omega = \Omega(P)$ denote the set of all linear extensions of P . A much-studied problem is the generation of a linear extension uniformly at random. This is similar to the problem of generating an EH of a sample of sequences uniformly at random due to the following. Given a sample of sequences that share a genealogical history, a perfect phylogeny can be constructed (efficiently [15]), which determines the evolutionary events (coalescences and back mutations, if viewed backwards in time) that happen along an EH. The EHs only differ in the order of these events, and thus every EH can be viewed as a permutation of the events. In this sense, the problem is similar to generating linear extensions of a poset. However, the restrictions on the order of the evolutionary events do not form a poset. For instance, consider the example in Figure 1 in Section 2. Here neither M_1 nor M_2 have to necessarily come before the first coalescent event; however, one of them has to. It is these kinds of restrictions that make the problem of generating EHs uniformly at random more complicated.

There has been much research on the problem of sampling from Ω uniformly and of approximating $|\Omega|$. Brightwell and Winkler [5] showed that determining $|\Omega|$ is $\sharp P$ -complete. Since the problem is self-reducible, a polynomial time algorithm for uniform sampling also yields a fully polynomial time approximation scheme (FPRAS) for $|\Omega|$.

The first such FPRAS came from Dyer, Frieze and Kannan [11], based on the rapid mixing of a particular geometric Markov chain, using the geometric bounds developed by Jerrum and Sinclair [20, 28]. Later, Karzanov and Khachiyan [21] showed the rapid mixing of a combinatorial Markov chain on Ω using, likewise, the geometric and conductance arguments of Jerrum and Sinclair [20, 28]. These results all rely on a relationship between Ω and a certain polytope in \mathbb{R}^n , for which strong isoperimetric inequalities hold. In the case of all possible EHs, however, the corresponding high-dimensional body is not convex, due to the nature of restrictions mentioned previously. Subsequently, the isoperimetric inequalities fail to hold and therefore the

methods cannot be transferred to this problem.

A non-geometric proof of rapid mixing of a slight modification of the Karzanov-Khachiyan chain was given by Bubley and Dyer [7] using the technique of path coupling [6]. When using coupling arguments, one has to define the coupling over all pairs of states. With path coupling, however, one defines a path between an arbitrary pair of states, and then only has to define the coupling over pairs of states that are adjacent in some path. If one can show that for all pairs of path-wise adjacent states, that the two Markov chains, with an appropriate coupling and metric, come closer together in expectation, then it follows that the entire path is contracting in expectation. The proof of rapid mixing then follows by induction and the Markov inequality.

Due to the nature of restrictions in the case of possible EHs, the proof presented in [7] cannot yet be transferred to the Markov chain defined in Section 3 in full generality. However, given a special case of samples of sequences, the proof can be transferred with a slight modification. We present the proof in the next section.

4.3 Example when path coupling works

Suppose our data set consists of $n + 1$ sequences, all of different type, all but one having one mutation:

$$S = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

In this case each EH consists of n back mutation events (which can be labeled M_1, \dots, M_n according to which segregating site the mutation is on), and n coalescence events of all 0 sequences. The back mutations can occur in an arbitrary order; we do, however, have one restriction: if viewed backwards in time, every prefix of an EH has to contain at least as many back mutations as coalescences. Let Ω denote the space of all EHs. From the above it follows that Ω is a product space: $\Omega = \Omega_1 \times \Omega_2$, where Ω_1 is the space of $2n$ -long sequences consisting of n M 's and n C 's and where every prefix contains at least as many M 's as C 's, and $\Omega_2 = S_n$, the symmetric group of order n . We know the size of the state space: the size of Ω_1 is the n^{th} Catalan number, $|\Omega_1| = C_n = \frac{1}{n+1} \binom{2n}{n}$, while $|\Omega_2| = n!$, and therefore $|\Omega| = |\Omega_1||\Omega_2| = \frac{(2n)!}{(n+1)!}$.

Thus we see that the number of possible EHs grows rapidly in n . Still, rapid mixing is possible to prove.

First of all, since the state space is a product space, it follows that the uniform distribution π on Ω is the product of uniform distributions π_1 and π_2 on Ω_1 and Ω_2 respectively. Therefore if there exists two rapidly mixing Markov chains, X_1 and X_2 on Ω_1 and Ω_2 respectively, then the “combing together” of the two, namely alternating between moves in X_1 and X_2 , will give a rapidly mixing Markov chain on Ω . Thus we consider Markov chains on Ω_1 and Ω_2 separately.

First, let us consider Ω_2 , the set of all permutations of a set of n elements. In this case the Markov chain described in Section 3 will mix rapidly, since it is essentially the Karzanov-Khachiyan chain, which is known to mix rapidly on the set of linear extensions of a poset—in this case there are no restrictions given by the partial order.

Now let us consider Ω_1 . We slightly modify the chain described in Section 3 according to [7]. Let f be a concave probability distribution on $\{1, 2, \dots, 2n - 1\}$. When choosing two consecutive events, instead of choosing according to the uniform distribution, we pick an $i \in \{1, 2, \dots, 2n - 1\}$ according to f , and choose events i and $i + 1$. Otherwise, the chain is unchanged. Note, however, that here we have unlabeled M 's and C 's; therefore if we pick two consecutive M 's or two consecutive C 's, then swapping them or not, we stay in the same state. When considering consecutive events where there is an M and a C as well, the only restriction that could prevent a swap from happening is that there must be at least as many M 's in every prefix as C 's. For notational convenience we shall assume that f is also defined on 0 and $2n$, and that $f(0) = f(2n) = 0$.

Bubley and Dyer's proof of rapid mixing using path coupling carries over, with a slight modification, to the current situation. We modify the definition of the distance between two states, causing only some cases in the proof to change. We write the proof out in detail in order to provide a clear understanding and also to highlight where it fails for the general case.

Similarly as in Section 3, let x and y be neighbors in Ω_1 if the swapping of two consecutive events can take x to y . For any $x, y \in \Omega_1$, let a path from x to y be any sequence z_0, z_1, \dots, z_r , where $x = z_0$ and $y = z_r$, such that $z_k \in \Omega_1$ and z_k and z_{k-1} are neighbors for $k = 1, 2, \dots, r$. Swapping events i_k and $i_k + 1$ takes z_{k-1} to z_k . Define the distance of two states in Ω_1 by the length of the shortest path between them. This is simply a graph distance, which we denote by δ . It is not hard to see that $D := \max_{x, y \in \Omega_1} \delta(x, y) \leq \sum_{i=1}^{n-1} i \leq n^2/2$.

Let X and Y be two copies of the Markov chain, as in the Coupling Lemma. At time t , let $X_t = Z_0, Z_1, \dots, Z_{d_t} = Y_t$ be a shortest path between

X_t and Y_t , with length d_t . Now we let the Z_k evolve for a single time step as coupled copies of the Markov chain. Let Z'_k denote the state to which Z_k evolves, and let d_{t+1} be the distance between X'_t and Y'_t . We couple the Z_k as follows.

1. Choose $p \in \{1, 2, \dots, 2n - 1\}$ according to the distribution f , and choose $c_0 \in \{0, 1\}$ uniformly at random.
2. For each $k \in \{1, 2, \dots, d_t\}$: if $p = i_k$, then let $c_k = 1 - c_{k-1}$, otherwise let $c_k = c_{k-1}$.
3. For each $k \in \{1, 2, \dots, d_t\}$: if $c_k = 0$ or the swapping of events p and $p + 1$ is not allowed in Z_k , then let $Z'_k = Z_k$; otherwise, let Z'_k be the EH that one gets by swapping events p and $p + 1$ in Z_k .

We will show that for this coupling

$$\mathbb{E}(\delta(Z'_{k-1}, Z'_k)) \leq 1 + \frac{1}{2}(f(i_k - 1) - 2f(i_k) + f(i_k + 1)) \quad (4.1)$$

holds for every k . Since $\delta(Z_{k-1}, Z_k) = 1$, it follows that the graph distance does not increase in expectation under this coupling for any concave f . We can choose f in such a way that the inequality above is a strict contraction, and from this we can show rapid mixing.

Choose the concave probability distribution to be $F(i) = i(2n - i)/K$, where K is the normalizing constant: $K = (4n^3 - n)/3$. Then it follows that for all i , $\frac{1}{2}(F(i - 1) - 2F(i) + F(i + 1)) = -1/K$. Now applying inequality 4.1, we have that for this Markov chain and this coupling

$$\begin{aligned} \mathbb{E}(d_{t+1}|d_t) &\leq \mathbb{E}\left(\sum_{k=1}^{d_t} \delta(Z'_{k-1}, Z'_k) | d_t\right) \\ &\leq \mathbb{E}\left(\sum_{k=1}^{d_t} (1 - 1/K) | d_t\right) = (1 - 1/K) d_t. \end{aligned}$$

Thus it follows that $\mathbb{E}(d_t) \leq (1 - 1/K)^t D$, and since d_t is a non-negative integer valued random variable, Markov's inequality gives us

$$\mathbb{P}(X_t \neq Y_t) = \mathbb{P}(d_t \geq 1) \leq \mathbb{E}(d_t) \leq (1 - 1/K)^t D.$$

Now applying the Coupling Lemma, we see that $d_{TV}(\mu_t, \pi) \leq (1 - 1/K)^t D$, where μ_t is the distribution of the Markov chain at time t and π is the

equilibrium distribution. Therefore in order to ensure that $d_{TV}(\mu_t, \pi) \leq \varepsilon$, it suffices to simulate the Markov chain for

$$K \log(D\varepsilon^{-1}) \leq \frac{1}{3}(4n^3 - n) \log\left(\frac{1}{2}n^2\varepsilon^{-1}\right) = O(n^3 \log(n\varepsilon^{-1}))$$

steps.

Thus to conclude the proof we have to show that 4.1 holds for every k . For notational simplicity, let us write $A = Z_{k-1}$, $B = Z_k$ and $i = i_k$. Thus swapping events i and $i + 1$ in A takes us to B . This means that A and B agree in every event except i and $i + 1$, and that events i and $i + 1$ are M and C respectively in one EH, and C and M respectively in the other. In the following we differentiate between different cases depending on the value of p , and then we use

$$\mathbb{E}(\delta(A', B')) = \sum_{j=1}^{2n-1} f(p) \mathbb{E}(\delta(A', B') | p = j). \quad (4.2)$$

If $p \notin \{i - 1, i, i + 1\}$ then $\delta(A', B') = \delta(A, B) = 1$, since either we do nothing in both A and B , or the swapping of events p and $p + 1$ can be applied in both A and B , and in this case A and B still agree in every event except i and $i + 1$, and thus are one swap away.

If $p = i - 1$ then with probability $1/2$ we do nothing in both A and B , and with probability $1/2$ we attempt to make the swap of events $i - 1$ and i in both A and B . If we do nothing in both A and B , then $\delta(A', B') = \delta(A, B) = 1$. If we attempt a move then the following observation is key, and this is where the proof fails to carry over in the general case: we know that only two kinds of events are possible, M 's or C 's, and we also know that A and B agree in event $i - 1$ but do not agree in event i . Therefore the events $i - 1$ and i must be of the same type for either A or B . For this EH, the swapping of these events does not change anything, and therefore only one swap can be relevant, so in this case $\delta(A', B') \leq 2$. (In the general case it might happen that $\delta(A', B') = 3$.) Consequently, $\mathbb{E}(\delta(A', B') | p = i - 1) \leq 3/2$. The $p = i + 1$ case is similar.

If $p = i$ then we can swap events i and $i + 1$ in both A and B , since this is how we get one EH from the other. Moreover, the coupling guarantees that we do nothing in one and we swap events i and $i + 1$ in the other. Therefore in this case $A' = B'$ with unit probability, and so $\mathbb{E}(\delta(A', B') | p = i) = 0$.

Combining the results from the various cases and using 4.2, we arrive at

$$\mathbb{E}(\delta(A', B')) = 1 + \frac{1}{2}(f(i - 1) - 2f(i) + f(i + 1)),$$

which concludes the proof.

4.4 Lower bound on the mixing rate

In this section we show an example (a particular data set) where the Markov chain described in the previous section has mixing time $\Omega(n^3)$ for any concave choice of f , and thus we have a lower bound of $\Omega(n^3)$ for the mixing time in the worst case.

Consider the data set that consists of $n+1$ sequences, only one segregating site, and only one sequence carrying a mutation in the segregating site:

$$S = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} n \\ 1 \end{pmatrix}.$$

For this data set there are n possible EHs, each consisting of n coalescent events and 1 mutation event: the mutation event can be any event except the last one (when viewed backward in time). The Markov chain described above is therefore equivalent to a random walk on $\{1, 2, \dots, n\}$, denoting where the mutation event is, and moving with probabilities given by f . Since f is a concave probability distribution, its maximum is at most $2/n$. Thus the expected time before the Markov chain performs a non-trivial transition is $\Omega(n)$. Since a simple random walk on $\{1, 2, \dots, n\}$ has mixing time $\theta(n^2)$, it follows that the Markov chain has a mixing time of $\Omega(n^3)$ on this data set.

5 Conclusions and Future Work

In this report, we investigated Markov chain Monte Carlo techniques that sample from evolutionary histories of a data set, assuming the infinite sites model. We constructed a Markov chain which has a uniform stationary distribution and is irreducible over the set of all evolutionary histories. While rapid mixing of this chain in the general case has not yet been shown, we succeeded in proving rapid mixing in a special case.

We wish to extend our results to a proof of rapid mixing in the general setting, and then consider the chain whose stationary distribution is not uniform, but the distribution given by the probabilities of the evolutionary histories. It will then be important to look at the biological implications of the results.

Acknowledgements.

I thank my advisors Jotun Hein, Rune Lyngsø and István Miklós for proposing this problem to work on, and for guidance and support throughout the

project. I also thank the Department of Statistics at the University of Oxford for the kind hospitality, and the PolyGene Project and the Bizáki Puky Péter Emlékalapítvány for funding.

A Acceptance Ratio Can Be Small

Since our goal is to sample from the EHs according to the probabilities described in Section 2.3 (these probabilities do not sum to one, but they can be normalized if we divide them by the probability of the data set (which is, unfortunately, unknown)), it is of interest to know how “far” the probabilities of different EHs can be from each other. A way of measuring this is by looking at the ratio of the probabilities of the most likely (i.e., the one with highest probability among possible EHs) and the most unlikely evolutionary histories. This ratio depends on n (the number of sequences in the data set), m (the number of segregating sites in the data set) and θ (the scaled mutation rate), but does not depend on the normalizing factor. The following example shows that this ratio can grow as $(c_1 n)^{c_2 M}$, where c_1 and c_2 are appropriate positive constants.

Consider a data set with only two types of sequences: all 0 sequences with multiplicity l and all 1 sequences with multiplicity $n - l$. The number of segregating sites is m , so there are m 1’s in the all 1 sequences. Consider the following two histories:

- *History A:* Going from the data set to the MRCA, first the all 0 sequences coalesce until their multiplicity is 1, then the all 1 sequences coalesce until their multiplicity is 1, then there are m back mutations (in an arbitrary order) and in the end there is a coalescence of the two all 0 sequences.
- *History B:* Going from the data set to the MRCA, first the all 1 sequences coalesce until their multiplicity is 1, then there are m back mutations (in an arbitrary order), and then the remaining $l + 1$ all 0 sequences coalesce until their multiplicity is 1, i.e. they reach the MRCA.

Let us denote the probabilities of the two histories $\pi(A)$ and $\pi(B)$ respectively. When calculating the ratio $\pi(A)/\pi(B)$ the normalizing constants cancel each other out, and therefore we only need to take into account the appropriate factors from the EGT recursion (as described in Section 2.3) when calculating the probabilities.

For history A , the first $l - 1$ events are coalescences of all 0 sequences. These give the following factors to $\pi(A)$:

$$\frac{l-1}{n-1+\theta} \frac{l-2}{n-2+\theta} \cdots \frac{1}{n-l+1+\theta}.$$

Then the next $n-l-1$ events are coalescences of all 1 sequences. These give the following factors to $\pi(A)$:

$$\frac{n-l-1}{n-l+\theta} \frac{n-l-2}{n-l-1+\theta} \cdots \frac{1}{2+\theta}.$$

Now there are two sequences left, an all 0 sequence and an all 1 sequence. The next m events are back mutations (in an arbitrary order). The first $m-1$ of these are of the same case: the type after the back mutation is different of all types present in the sample (in this case the only type is the all 0 sequence); however, in the case of the last back mutation we arrive at an all 0 sequence, which is already present in our sample, and thus the appropriate factor from the EGT recursion is of the other case. Altogether the factors from the mutations to $\pi(A)$ are the following:

$$\left(\frac{\theta}{2(1+\theta)} \right)^{m-1} \frac{2\theta}{2(1+\theta)}.$$

Finally, the last event is a coalescence of two all 0 sequences, giving a factor of

$$\frac{1}{1+\theta}.$$

For history B , the first $n-l-1$ events are coalescences of all 1 sequences. These give the following factors to $\pi(B)$:

$$\frac{n-l-1}{n-1+\theta} \frac{n-l-2}{n-2+\theta} \cdots \frac{1}{l+1+\theta}.$$

Then the next m events are back mutations (in an arbitrary order). Similarly as before, the first $m-1$ of these are of the same case: the type after the back mutation is different of all types present in the sample (in this case the only type is the all 0 sequence); however, in the case of the last back mutation we arrive at an all 0 sequence, which is already present in our sample, and thus the appropriate factor from the EGT recursion is of the other case. The factors from the mutations to $\pi(B)$ are the following:

$$\left(\frac{\theta}{(l+1)(l+\theta)} \right)^{m-1} \frac{(l+1)\theta}{(l+1)(l+\theta)}.$$

Finally, the last l events are coalescences of all 0 sequences, giving the following factors to $\pi(B)$:

$$\frac{l}{l+\theta} \frac{l-1}{l-1+\theta} \cdots \frac{1}{1+\theta}.$$

When taking the ratio $\pi(A)/\pi(B)$, there are a lot of cancellations and what remains is the following:

$$\frac{\pi(A)}{\pi(B)} = \frac{1}{l} \left(\frac{(l+1)(l+\theta)}{2(1+\theta)} \right)^{m-1} \frac{l+\theta}{1+\theta}.$$

For $\theta = 1$ this gives us:

$$\frac{\pi(A)}{\pi(B)} = \frac{1}{l} \left(\frac{(l+1)}{2} \right)^{2m-1}.$$

So for $\theta = 1$ and $l = n/2$ we arrive at

$$\frac{\pi(A)}{\pi(B)} = \frac{2}{n} \left(\frac{\left(\frac{n}{2} + 1\right)}{2} \right)^{2m-1} \geq \frac{1}{2} \left(\frac{n}{4} \right)^{2m-2},$$

and for $\theta = 1$ and $l = n - 1$ we arrive at

$$\frac{\pi(A)}{\pi(B)} = \frac{1}{n-1} \left(\frac{n}{2} \right)^{2m-1} \geq \frac{1}{2} \left(\frac{n}{2} \right)^{2m-2}.$$

References

- [1] D. Aldous. Random walks on finite groups and rapidly mixing Markov chains. *Séminaire de Probabilités*, XVII, 1981-1982.
- [2] D. Aldous and P. Diaconis. Shuffling Cards and Stopping Times. *American Mathematical Monthly*, pages 333–348, 1986.
- [3] D. Aldous and J. Fill. Reversible Markov Chains and Random Walks on Graphs. Book in preparation, available via <http://www.stat.berkeley.edu/~aldous/RWG/book.html>.
- [4] C.M. Andreassen and A. Okholm. From exact marginals to better importance sampling. Technical report, University of Oxford, 2008.
- [5] G. Brightwell and P. Winkler. Counting Linear Extensions. *Order*, 8(3):225–242, 1991.

- [6] R. Bubley and M. Dyer. Path Coupling: A Technique for Proving Rapid Mixing in Markov Chains. In *Proceedings of the Thirty-Eighth Annual Symposium on Foundations of Computer Science*, pages 223–231, 1997.
- [7] R. Bubley and M. Dyer. Faster Random Generation of Linear Extensions. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 350–354. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 1998.
- [8] P. Diaconis. Group Representations in Probability and Statistics. *Lecture Notes Monograph Series Vol. 11, Institute of Mathematical Sciences, Hayward, California*, 1988.
- [9] P. Diaconis and D. Stroock. Geometric Bounds for Eigenvalues of Markov Chains. *The Annals of Applied Probability*, pages 36–61, 1991.
- [10] W. Doeblin. Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états. *Rev. Math. Union Interbalkanique*, 2:77–105, 1938.
- [11] M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the Association for Computing Machinery*, 38(1):1–17, 1991.
- [12] S.N. Ethier and R.C. Griffiths. The infinitely-many-sites model as a measure-valued diffusion. *The Annals of Probability*, pages 515–545, 1987.
- [13] J. Felsenstein, M. Kuhner, J. Yamato, and P. Beerli. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology*, page 163. Institute of Mathematical Sciences, 1999.
- [14] R.C. Griffiths and S. Tavaré. Unrooted Genealogical Tree Probabilities in the Infinitely-Many-Sites Model. *Mathematical Biosciences*, 127(1):77–98, 1995.
- [15] D. Gusfield. Efficient Algorithms for Inferring Evolutionary Trees. *Networks*, 21(1):19–28, 1991.
- [16] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

- [17] J. Hein, M.H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, 2005.
- [18] A. Hobolth, M.K. Uyenoyama, and C. Wiuf. Importance Sampling for the Infinite Sites Model. *Statistical Applications in Genetics and Molecular Biology*, 7(1):32, 2008.
- [19] F. Huszar and S.G. O’Keeffe. "Corner-Cutting" Approaches to the Ethier-Griffiths-Tavare Recursions. Technical report, University of Oxford, 2008.
- [20] M. Jerrum and A. Sinclair. Approximating the Permanent. *SIAM Journal on Computing*, 18:1149, 1989.
- [21] A. Karzanov and L. Khachiyan. On the Conductance of Order Markov Chains. *Order*, 8(1):7–15, 1991.
- [22] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, 1969.
- [23] J.F.C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.
- [24] M.K. Kuhner, J. Yamato, and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140(4):1421–1430, 1995.
- [25] J.S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.
- [26] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21(6):1087, 1953.
- [27] A. Sinclair. Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, Probability and Computing*, 1(04):351–370, 1992.
- [28] A. Sinclair and M. Jerrum. Approximate Counting, Uniform Generation and Rapidly Mixing Markov Chains. *Information and Computation*, 82(1):93–133, 1989.

- [29] M. Stephens and P. Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 605–655, 2000.
- [30] I.J. Wilson and D.J. Balding. Genealogical inference from microsatellite data. *Genetics*, 150(1):499–510, 1998.
- [31] I.J. Wilson, M.E. Weale, and D.J. Balding. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 155–201, 2003.