Modeling Gene-Tree-Species-Tree Conflict with Migration Using Continuous-Time Markov Chains

Haleigh Miller

Advisors: Bruce Rannala, Elbridge Gerry Puckett

March 20, 2020

Abstract

From virology to conservation biology, phylogenetic trees are ubiquitous across scientific disciplines. To infer a phylogenetic species tree from genetic data a researcher encounters a common question: "what software tool should I use?" The choice is often between a summary-based heuristic tool or a computationally intensive but statistically consistent Bayesian inference tool. We modeled migration and coalescence in a symmetric 4-population tree using a continuous-time Markov chain to discover scenarios under which summary-based methods will produce erroneous estimates of phylogeny whereas Bayesian inference methods based on the correct model will not. We present a novel algorithm for constructing the generator matrix of the Markov chain, which can be used to calculate the probability of gene-tree-species-tree conflict and explore regions of parameter space under which conflict is more probable and thus erroneous estimates of phylogeny by heuristic estimators are more probable.

Contents

1 Introduction

1.1 Background

Molecular phylogenetics is the study of evolutionary relationships between a group of organisms using molecular data, such as DNA or amino acid sequences. Since the inception of evolutionary theory, scientists have been interested in the tree of life, how all species are related to each other and the story of their ancestors and divergences. Originally, morphological characters were used to construct phylogenies. However, convergent evolution can make morphological information unreliable. In 1962, Emile Zuckerkandl and Linus Pauling published *Molecules as Documents of Evolutionary History* and launched the field of molecular

phylogenetics [26]. The crux of the field is that similarities in homologous DNA sequences are informative of evolutionary relatedness.

Formally, a phylogenetic tree is a hypothesis of the genealogical relationships among species, among genes, among populations and even among individuals. Here, we specifically consider a tree of different populations, but often species and population trees are used interchangeably. The population tree is a directed graph where the external nodes, called **tips**, represent present-day populations and internal nodes represent ancestral populations. Given a tree with n tips, there are $2n-1$ nodes and $2n-2$ edges. Forward in time, branches descended from an internal node (ancestral population) represent a **population divergence event**. Figure 1 below demonstrates the terminology.

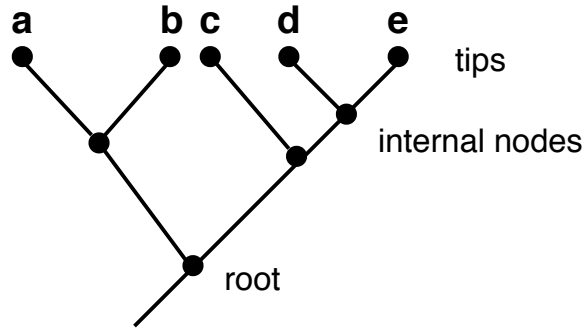


Figure 1: Tree Terminology

Phylogenetic trees are often inferred from DNA nucleotide sequence data, which is called **phylogenetic inference**. We can align a region of genome, called a genetic **locus**, from our organisms of interest and examine the number of differences, from mutations or insertions, between them through the use of a sequence-distance measure. There is assumed to be an ancestral version of the locus from which all the present-day sequences are derived. Pairwise distances between loci from different species are used to build a gene tree for the given locus.

DNA acts as a window into the past with which we can reconstruct the ancient history of populations. But, DNA tells more than one story; different areas of the genome may have different evolutionary histories and can result in different species trees [14, 19, 6]. Inferring

the evolutionary history of a group of organisms given discordant genomic histories is a central problem in phylogenetics. There are many reasons for the evolutionary history of a particular gene, called a **gene tree**, to have a different evolutionary history than the organisms it comes from including the coalescent process in ancestral species, gene duplication, and horizontal gene transfer (introgression) [14, 19, 6]. In fact, there are situations under which the most probable gene tree has a different topology (branching relationships) than the species tree, which is called the "anomaly zone" [4]. The coalescent process in ancestral species that generates conflicts between gene tree ans the species tree is often referred to as **incomplete lineage sorting (ILS)**. Gene tree heterogeneity due to the coalescent process is universal [7] and the influence it has on gene-tree-species-tree conflict is significant. In fact, simulation studies have shown that the probability that two gene trees match can be very small in the presence of ILS [22]. Because of the pervasiveness of ILS and its potentially mileading effects, many phylogenetic inference programs are designed to incorporate this process. The **multispecies coalescent (MSC)** has emerged as the natural framework to account for genealogical heterogeneity due to ILS across the autosomal genome [7].

1.2 Coalescent-Based Phylogenetic Inference

1.2.1 Coalescent Theory

The MSC is an extension of the coalescent theory formulated by J.F.C. Kingman in the 1980s [11]. The coalescent model traces ancestral lineages of a sample of chromosomes from a population backwards in time. An ancestral lineage is a series of genetic ancestors of the sample at a locus. With each coalescent event, the number of ancestors decreases by 1. In the coalescent model, we are concerned with the time until we have $i-1$ ancestors, denoted T_i. When $i = 1$, we have reached the the **most recent common ancestor** and T_2 is denoted T_{MRCA} [23].

In the basic coalescent model we make the following assumptions:

1. No selection, namely no differences in fitness (reproductive success) among lineages at a locus.

2. There is no population subdivision, geographic or otherwise.

3. The size of the population, N, is constant through time.

Assumptions 1 and 2 mean that the number of offspring each individual in the sample produces is independent of any label that could be placed on them, be that label allelic state, geographic location, etc. This property is called **exchangeability**. Exchangeability implies that every pair of lineages is equally likely to be the pair that coalesces, a fact that will be useful later in the derivation of our Markov chain.

The coalescent is a stochastic process in which the state space is all possible rooted trees with ordered tips and nodes and associated coalescent times, referred to as a **labeled histories** [5] . Since every pair of lineages is equally likely to coalesce, all labeled histories in the state space have equal probability. Kingman [11] showed that as $N \to \infty$ the **waiting times** between coalescence events $t_i = T_i - T_{i-1}$ are independent and identically distributed (i.i.d) exponential random variables, i.e., that each coalescence time has the probability density function,

$$f(t_i) = \binom{i}{2} e^{-\binom{i}{2} t_i} \quad \forall i = 1, ..., n-1 \tag{1}$$

From the properties of exponential distributions it is clear that,

$$\mathbb{E}[t_i] = \frac{2}{i(i-1)} \tag{2}$$

and

$$Var[t_i] = \left(\frac{2}{i(i-1)}\right)^2 \tag{3}$$

1.2.2 Multispecies Coalescent Model

The MSC lies at the interface of population genetics and phylogenetics [20]. The MSC extends coalescent theory by accounting for the history of species and population divergences. It extends traditional phylogenetic models by accounting for the coalescent process and the resulting genealogical heterogeneity across the genome. Because it accounts for the coalescent process in both extant and extinct ancestral species, the MSC naturally accommodates ILS.

The MSC has two types of parameters: **species divergence times** (τs) and **population sizes** (θs) [7]. Both τs and θs are estimated by sequence distance, the expected number of mutations/substitutions per site. The parameter $\theta = 4N\mu$ is the average distance between two sequences sampled at random from a population with effective population size N, where μ is the mutation rate per site per generation. For example, in humans $\theta = 0.0006$; meaning, if you compared two sequences in humans you would expect 0.6 differences between them per 1000 base pairs. Divergence times in a population tree are also represented in units of expected substitution. For example,τ represents the age of an internal node in the species tree, measured in units of expected number of mutations per site. Figure 2 shows the relationship betweeen parameters of gene trees and population trees for a population tree of 3 species.

Phylogenetic inference methods built on the MSC, often called coalescent-based methods, fall into two categories: summary and full-likelihood methods. In full-likelihood MSC methods, a statistical methodology is needed to infer the probability of a species tree given the gene trees. Bayesian inference of species trees was first developed by Ziheng Yang and Bruce Rannala in the mid 1990's [21]. Bayesian statistics are useful for this problem because prior probability distributions can be used to describe the uncertainty of all unknowns, including the model parameters. Let $\mathbf{\Theta}$ represent the parameters of the MSC and the DNA substitution model and D be the observed data, i.e., the sequence data. Bayes' theorem

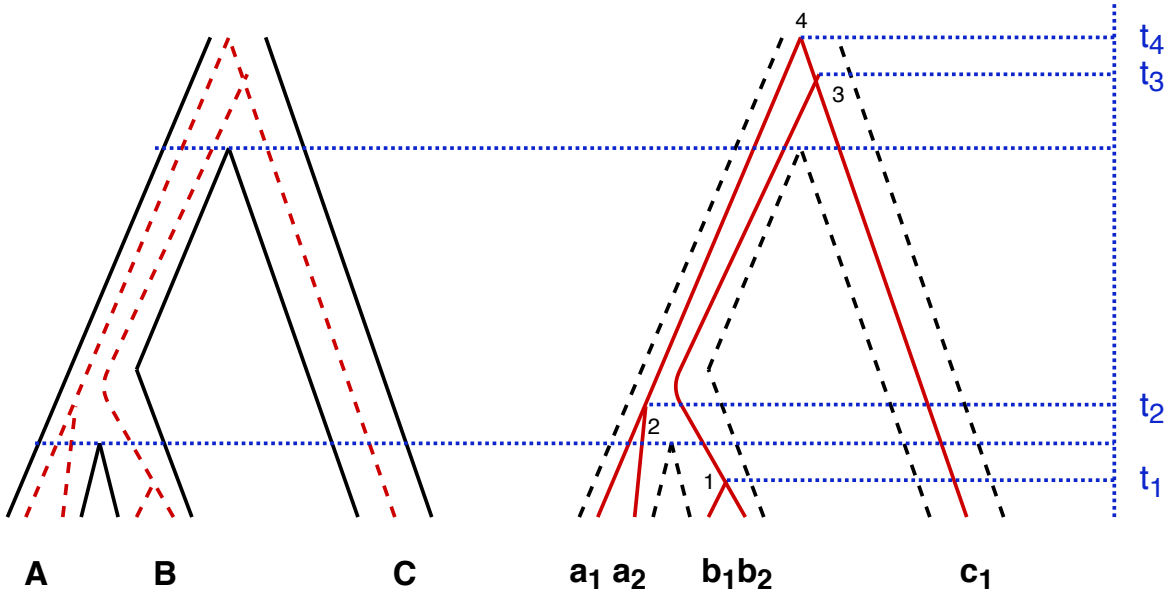


Figure 2: A population tree for populations A, B and C is shown in black with a gene tree for sequences a_1, a_2, b_1, b_2, and b_3 shown within the species tree in red. Within each species/population, sequences coalesce at random at the rate determined by θ, generating a gene tree with branch lengths, t_is, conditioned on the species tree. Note that θ_c not estimable if there is only one sequence from species C at each locus.

states that,

$$f(\mathbf{\Theta}|D) = \frac{1}{z} f(\mathbf{\Theta}) f(D|\mathbf{\Theta})$$

where z is the normalizing constant, $z = \int f(\mathbf{\Theta}) f(D|\mathbf{\Theta})$, and $f(\mathbf{\Theta})$ is the prior distribution of the parameters.

Bayes' theorem implies that the posterior is proportional to the prior multiplied by the likelihood, i.e., that the posterior combines information about the likelihood and the prior. With respect to the MSC, we wish to find the posterior probability of a species tree given the model parameters, the posterior probability of the associated gene tree, and the sequence alignment. Instead of calculating the posterior, which is computationally unfeasible because of the integral calculation in the normalizing constant, a Markov chain Monte Carlo (MCMC)

algorithm obtains a sample from the posterior.

The MSC is, however, computationally cumbersome with large datasets; the more species or loci added, the more complicated the model becomes. Heuristic methods, or **summary-statistic methods**, make the method more computationally efficient by treating the gene tree as static data, or as a summary-statistic. But, by making the gene trees static, the uncertainty in each gene tree is ignored resulting in a potential loss of **statistical efficiency and consistency**, the assurance that with more loci added the method will have minimal variance and converge to the correct tree.

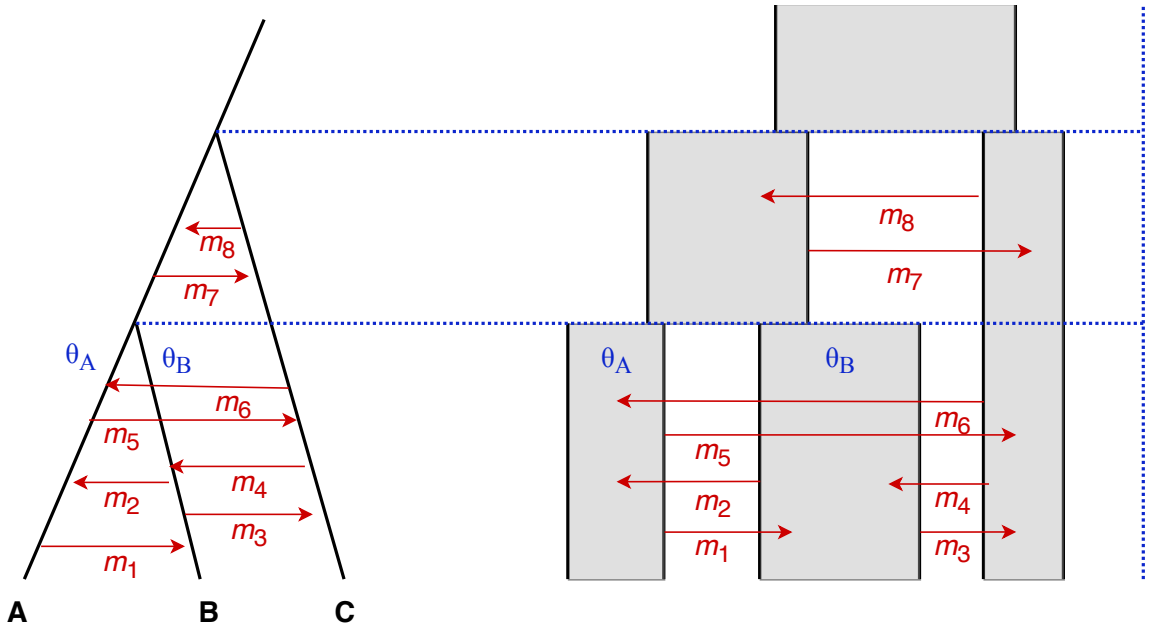
MP-EST [12] and ASTRAL [17, 16, 24], two popular summary-statistic methods, are based on breaking a species tree into 3- or 4-taxa sub-trees, respectively. This is based on the findings that there are no anomalous rooted 3-population species trees [4] and no anomalous rooted 4-population species trees [1, 3]. These methods find the species tree topology that agrees with the largest number of sub-tree topologies induced by the inferred gene trees. However, if the most common topology of the gene trees is discordant with the species tree topology, these methods will be inconsistent.

A major question in the field is whether the increased performance of rigorous Bayes inference methods, which are not subject to the inconsistencies of summary-statistic methods, justifies the increased computational cost. Further, since both classes of methods derive from the MSC, they share the following assumptions: there is (i) no selection on the genes, (ii) no migration between populations, and (iii) no recombination within loci. This raises another question: is one class of methods more resilient to violations of model assumptions, such as migration, than the other? Moreover, is it necessary to develop Bayesian methods that model migration to obtain accurate phylogenetic inferences?

1.2.3 Multispecies Coalescent Model with Migration

Migration has been shown to significantly influence dynamics of the divergence process between populations or species [8]. To model migration in a MSC process, in order to un-

derstand its influence, researchers often use the **Isolation-with-Migration (IM) Model**, which was first implemented by Jody Hey and Rasmus Nielsen in 2004 for two populations [9] and expanded to an arbitrary number of populations by Jody Hey in 2010 [8]. The IM model, shown in Figure 3, includes three population size parameters (θs) and two migration rate parameters (m_is) per divergence event. Migration rates are defined as $m_i = \frac{M}{\mu}$ where M is the migration rate per generation per copy and μ is mutation rate per site per generation. Thus, the migration rate is scaled in time units of the expected number of mutations. It is assumed that all ancestral and sampled populations are constant in size and follow Fisher-Wright assumptions[1]. Migration happens throughout a time period randomly at rate m_i. It is further assumed the loci sampled, like in the MSC, have not undergone recombination, are unlinked, and are not under selection.



00

Figure 3: IM model in 3-population asymmetric species tree. On the left, the IM model is illustrated in context of the species tree. On the right, the IM model is illustrated in context of the populations undergoing gene flow. Individuals in each population migrate at random at the rate determined by m_is

[1]Under Wright-Fisher assumptions there is no selection, no mutation, non-overlapping generation times and random mating.

The effect of migration on phylogenetic inference has been well-studied in 3-population IM models. In 2012, Tianqi Zhu and Ziheng Yang [25] derived the probability of gene-tree-species-tree discordance when migration is present between sister populations (extant populations that share a common ancestor not common to any other populations) in a 3-population tree. Using their 3-population IM model, they derived a likelihood-ratio test of speciation with gene-flow. In 2018, Colby Long and Laura Kubatko [13] derived the probability of gene-tree-species-tree discordance in a 3-population IM model with migration between all populations. They further described the **gene tree anomaly zone** as regions of parameter space in which discordant gene trees have higher probability than concordant ones. When migration rates were asymmetric, the probability of discordant trees was much higher than under symmetric migration rates. Analytic results such as these are useful because they can inform researchers of how to choose parameters in simulation studies of phylogenetic inference under migration.

Using a continuous-time Markov chain, we modeled migration and coalescence in a symmetric 4-population IM model. We derive a novel algorithm for automatically constructing the generator matrix that generates the Markov chain. From the generator matrix we can derive the probability of gene-tree-species-tree discordance and describe a gene tree anomaly zone for this model.

1.3 Continuous-Time Markov Chains

We model migration and coalescence in a phylogenetic tree using continuous-time Markov chains. Markov chains allow us to model the coalescent and migration history of sampled chromosomes of a locus backwards in time. Since the coalescent history of the sampled locus determines gene tree topology, we can calculate the probability of certain gene tree topologies given coalescent and migration parameters. We will outline the general theory of continuous-time Markov chains [2] before we define the chain for or coalescent with migration model.

Definition 1. *Let $\mathcal{F}_{X(s)}$ denote all the information pertaining to the history of a random variable X up to time s. Suppose $j \in S$ and $s \leq t$ and S is the countable state space of X. A process that satisfies the following conditions,*

$$P\{X(t) = j|\mathcal{F}_{X(s)}\} = P\{X(t) = j|X(s)\} \tag{4}$$

$$P\{X(t) = j|X(s)\} = P\{X(t-s) = j|X(0)\}, \tag{5}$$

*is called a **continuous-time Markov chain (CTMC)**. In other words, a CTMC is a process that satisfies both lack-of-memory and time-homogeneity properties.*

We are interested in the probability of reaching a certain state, say state j, of the CTMC given a starting state, say state i, in a finite time period t. In other words, we are interested in finding,

$$P_{ij}(t) = P(X(t) = j|X(0) = i) \tag{6}$$

where $X(t)$ is the state of the CTMC at time t.

Let $\lambda(i)$ denote the rate at which the process leaves state i and p_{ij} denote the probability that the process transitions to state j after leaving state i. $\lambda(i,j) = \lambda(i)p_{ij}$ represents the local rate of transitioning from state i to state j, also called the **transition intensity**. The leaving rate for state i is the sum of the local rates out of i, i.e.,

$$\sum_{j \neq i} \lambda(i,j) = \lambda(i)$$

These transition intensities are used to build the generator matrix which fully determines the chain. More formally,

Definition 2. *Let X(t) be a CTMC on some state pace S with transition intensities $\lambda(i,j) \geq$*

0 . *Then the matrix,*

$$A_{ij} = \begin{cases} -\lambda(i) & i = j \\ \lambda(i,j) & i \neq j \end{cases}$$

is called the ***generator*** *of the CTMC.*

Equation (3) is solved by the **Kolmogorov forward equations**:

$$P_i'(t) = -\lambda(j)P_{ij}(t) + \sum_{j \neq i} P_{ij}(t)\lambda(i,j) \tag{7}$$

From (4) and the definition of the generator, one can see that (3) can be written as the matrix differential equation,

$$P'(t) = P(t)A \tag{8}$$

The solution to (8) has the general form,

$$P(t) = e^{At} \tag{9}$$

Exponentiation of A analytically may work for certain small matrices, but for larger matrices numerical methods must be used. While there are many methods to compute the matrix exponential, issues with computational stability and efficiency are of concern in all [18]. The scaling-and-squaring method is the most widely used method for computing the matrix exponential [18], with current implementations having nearly optimal efficiency [10].

2 Methods

2.1 Model Description

To explore situations under which migration can cause gene-tree-species-tree conflict, and thus erroneous ASTRAL or MP-EST estimates of phylogeny, we have developed a model to allow for 4 populations in a fully symmetrical species tree, shown in Figure 4. We sample one sequence per species.

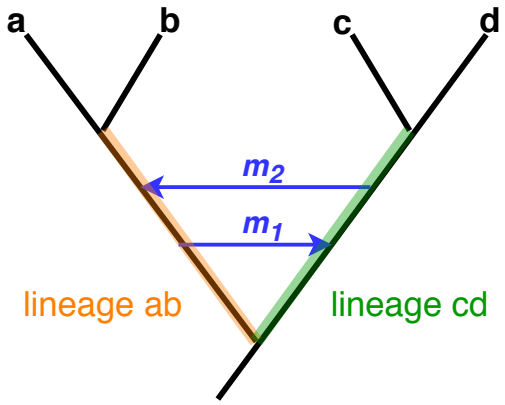


Figure 4: IM Model for a Symmetric Four-Population Tree. Migration occurs between ancestral populations ab and cd randomly at rates m_1 and m_2.

We construct a CTMC $X(t)$ in which we follow a single sequence from each of the populations a, b, c, and d in the 4-population phylogenetic tree backwards in time. Either a sequence can coalesce with another sequence or it can migrate to another population. Each element in the state space of $X(t)$ contains two sets, L and R. Let $L = \{\}$ denote the set of sequences in ancestral population ab and $R = \{\}$ denote the set of sequences in ancestral population cd. Let the set elements 1,2,3,4 denote sequences from populations a, b, c, d, respectively. So, for example, state $L = \{1, 2, 3\}R = \{4\}$ signifies that sequences from populations a, b and c are in population ab while the sequence from population d is in population cd. $L = \{1, 2, 3\}R = \{4\}$ will be represented as $\{1, 2, 3\}\{4\}$ for brevity. Let coalescence be denoted by the concatenation of set elements. For instance, if sequences 1

and 2 coalesce the coalesced lineage will be denoted 12.

2.2 Constructing the Generator Matrix

We assume that only one event, either a migration or a coalescence, can happen at any instance in time. To derive the coalescent event transition intensities we must return to coalescent theory; ancestors are chosen at random from the pool of $2N$ genes in a population of N individuals. Since the property of exchangeability holds, coalescence between any pair of sequences in the population is equally likely. If i sequences exist, the probability that a coalescence occurs between a random pair of sequences at any instance in time is given by,

$$P(\text{Coalescence}) = \binom{i}{2}\frac{1}{2N}$$

which is called the **instantaneous rate of coalescence**. However, the probability that two specified sequences coalesce in an instance in time within a population is just $\frac{1}{2N}$, which is our transition intensity for a viable coalescence event. The transition intensity for a migration event is simply the rate of migration in the appropriate direction.

The state space for this model has 92 states, which is too large to solve by hand, so we design an algorithm to build the generator matrix. Let $A = [a_{ij}] \in \mathbb{R}^{92\times 92}$ be our generator matrix, where $\boldsymbol{A}_{i,:}$ are the rows of A and $\boldsymbol{A}_{:,j}$ are the columns of A. We define row labels as **source** states and column labels as **target** states. So, a_{ij} is the transition intensity from source i to target j. We denote the row and column labels as $\boldsymbol{R_{i,:}}$.label–so if the ith row of A corresponds to source state $\{1\}\{2,3,4\}$, then $A_{i,:}$.label $= \{1\}\{2,3,4\}$. Further, L_i and R_i are the sets from population ab and population cd, respectively, in the source state. Likewise, L_j and R_j are the sets from population ab and population cd, respectively, in the target state.

Now, to assemble the matrix we define the following distance measures:

Definition 3. *The **state distance** between states L_iR_i and L_jR_j, denoted Δ^1_{ij}, is defined as,*

$$\Delta_{ij}^1 = |L_i \cup R_i| - |L_j \cup R_j|.$$

It represents the change in the number of sequences from the source state to the target state. When a migration event happens $\Delta_{ij}^1 = 0$ *and when a coalescent event happens* $\Delta_{ij}^1 = 1$*. All other values of* Δ_{ij}^1 *indicate there is no possible transition between the source and target states.*

Definition 4. *The* ***migration distance*** *between states* L_iR_i *and* L_jR_j*, denoted* Δ_{ij}^2*, is defined as,*

$$\Delta_{ij}^2 = |L_j - L_i| - |L_i - L_j|.$$

It represents the direction and magnitude of migration between two states. We only allow one sequence to migrate at a time so $|\Delta_{ij}^2| = 1$ *indicates a potential migration event.* $\Delta_{ij}^2 < 0$ *corresponds to migration from population 2 to population 1, while* $\Delta_{ij}^2 > 0$ *corresponds to migration from population 1 to population 2*

Definition 5. *The* ***coalescent distance*** *between states* L_iR_i *and* L_jR_j*, denoted* Δ_{ij}^3*, is defined as,*

$$\Delta_{ij}^3 = (|L_i| - |L_j|) - (|R_i| - |R_j|).$$

It represents the number of coalescent events taking place and under which population they occur. Since only two sequences can coalesce at any time, $|\Delta_{ij}^3| = 2$*.* $\Delta_{ij}^3 < 0$ *indicates a coalescent event occurred in population 2, while* $\Delta_{ij}^3 > 0$ *indicates a coalescent event occured in population 1.*

3 Results

Our methodology resulted in the following algorithm which can be used in further studies. The pseudo-code for the algorithm for generating R is as follows:

Algorithm 1: Generating the Generator Matrix R

Input : m_1, m_2, N
Output: R
/* Loop through source states (rows) */
for $i = 0, ..., 81$ **do**
  $A_{i,:}$.label $= L_i R_i$
  /* Loop through target states (columns) */
  for $j = 0, ..., 81$ & $i \neq j$ **do**
    $A_{:,j}$.label = target $:= L_j R_j$
    /* possible migration states: */
    if $\Delta_1^{ij} == 0$ **then**
      /* $|L_i \cap L_j| = |L_i|$ checks if the elements in L_i are shared with L_j (except the new member from population 2). This, with $\Delta_{ij}^2 == -1$ means there was a migration event from population 2 to population 1. */
      if $(\Delta_{ij}^2 == -1)$ & $(|L_i \cap L_j| == |L_i|)$ **then**
        $a_{ij} = m_2$
      end
      /* If $|L_i \cap L_j| = |L_i| - 1$ then $|R_i \cap |R_j| = |R_i|$, which checks if the elements in R_i are shared with R_j (except the new member from population 2). This with $\Delta_{ij}^2 == 1$ means there was a migration event from population 1 to population 2. */
      if $(\Delta_{ij}^2 == 1)$ & $(|L_i \cap L_j| == |L_i| - 1)$ **then**
        $a_{ij} = m_1$
      end
      /* No migration */
      else
        $a_i j = 0$
      end
    end
    /* possible coalescent states: */
    if $\Delta_1^{ij} == 1$ **then**
      $\Delta_{ij}^3 = (|L_i - L_j| - |R_i - R_j|)$
      /* x measures the number of shared elements in both sets. If x= $(L_j \cup R_j) - 1$ then all states except the coalescent state match */
      $x = |L_i \cap L_j| + |R_i \cap R_j|$
      /* Final coalescent state */
      if $(|L_i| == 2$ & $|R_i| == 0)$ *or* $(|L_i| == 0$ & $|R_i| == 2)$ **then**
        $a_i j = \frac{1}{2N}$
      end
      /* Coalescent event in 1 */
      if $(\Delta_{ij}^3 == 2)$ & $(x == |L_j \cup R_j| - 1)$ **then**
        $a_{ij} = \frac{1}{2N}$
      end
      /* Coalescent event in 2 */
      if $(\Delta_{ij}^3 == -2)$ & $(x == |L_j \cup R_j| - 1)$ **then**
        $a_{ij} = \frac{1}{2N}$
      end
    end
  end
  /* Fill in the diagonal so that $\sum_{i=0}^{81} a_i j = 0$ */
  $a_{ii} = -\sum_{j=0|i\neq j}^{81} a_i j$;
end
return A

To illustrate the construction of the generator matrix of the system, I have taken a 9 state subset of the state-space, shown in Figure 5. This subset, along with the subset of all states possible through migration, was also used to test my generator-matrix-generating algorithm.

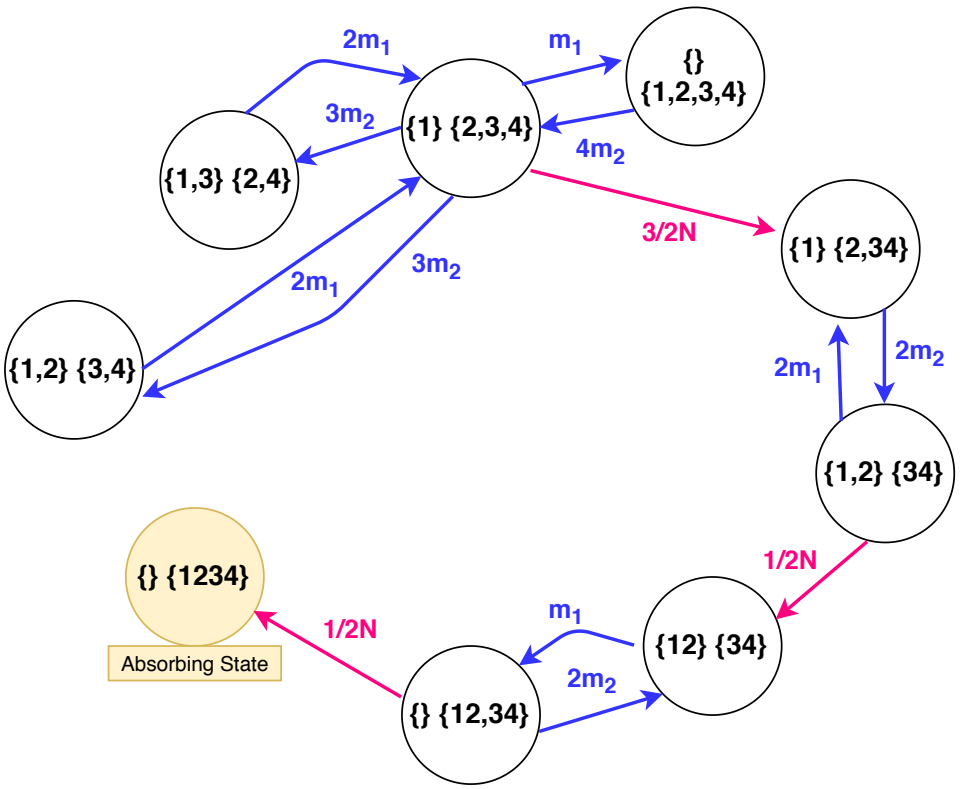


Figure 5: Illustration of Continuous Markov Chain for 4-Population IM Model on Subset of State-Space.

From Figure 5, R was found to be,

$$R = \begin{array}{c} \\ \{1,2\}\{3,4\} \\ \{1,3\}\{2,4\} \\ \{1\}\{2,3,4\} \\ \{\}\{1,2,3,4\} \\ \{1\}\{2,34\} \\ \{1,2\}\{34\} \\ \{12\}\{34\} \\ \{\}\{12,34\} \\ \{\}\{1234\} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \{1,2\}\{3,4\} & \{1,3\}\{2,4\} & \{1\}\{2,3,4\} & \{\}\{1,2,3,4\} & \{1\}\{2,34\} & \{1,2\}\{34\} & \{12\}\{34\} & \{\}\{12,34\} & \{\}\{1234\} \end{array} \\ \begin{bmatrix} -(2m_1\frac{1}{2N}) & 0 & 2m_1 & 0 & 0 & \frac{1}{2N} & 0 & 0 & 0 \\ 0 & -2m_1 & 2m_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3m_2 & 3m_2 & -(6m_2+m_1+\frac{3}{2N}) & m_1 & \frac{3}{2N} & 0 & 0 & 0 & 0 \\ 0 & 0 & 4m_2 & -4m_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2m_2 & 2m_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2m_1 & -(2m_1+\frac{1}{2N}) & \frac{1}{2N} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -m_1 & m_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2m_2 & -(2m_2+\frac{1}{2N}) & \frac{1}{2N} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

With $m_1 = 1$, $m_2 = 2$, $N = 400$ as sample values, using Algorithm 1 implemented in python, the generator matrix for the 9-state subset is below.

$$R = \begin{array}{c} \\ \{1,2\}\{3,4\} \\ \{1,3\}\{2,4\} \\ \{1\}\{2,3,4\} \\ \{\}\{1,2,3,4\} \\ \{1\}\{2,34\} \\ \{1,2\}\{34\} \\ \{12\}\{34\} \\ \{\}\{12,34\} \\ \{\}\{1234\} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \{1,2\}\{3,4\} & \{1,3\}\{2,4\} & \{1\}\{2,3,4\} & \{\}\{1,2,3,4\} & \{1\}\{2,34\} & \{1,2\}\{34\} & \{12\}\{34\} & \{\}\{12,34\} & \{\}\{1234\} \end{array} \\ \begin{bmatrix} -2.00125 & 0 & 2 & 0 & 0 & 0.00125 & 0 & 0 & 0 \\ 0 & -2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 6 & 6 & -13.0038 & 1 & 0.00375 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & -8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -4 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2.0025 & 0.00125 & 0.00125 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & -4.00125 & 0.00125 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0 \end{bmatrix} \end{array}$$

With this matrix, one can calculate the probability of discordance by calculating the probability of being in an anomalous state, such as $\{13\}\{23\}$, in varying times t. Recall that $P(t) = e^{At}$.The matrix exponential will be calculated using the scaling-and-squaring method [10] as implemented in *expm* in TensorFlow [15].

4 Conclusion

The motivation for this thesis was to develop analytical tools to study evolutionary scenarios under which popular phylogenetic inference methods were likely to infer incorrect phylogenies. This is an important problem in the field because, in choosing between a computationally-intensive but consistent Bayesian inference method or a computationally-efficient heuristic method, researchers must juggle the cost of computation against the danger of inconsistency. If there are only a few evolutionary scenarios under which heuristic methods were more likely to infer incorrect phylogenies, then it may not be necessary to use a Bayesian inference method.

We developed a probability model of a 4-population phylogenetic tree with migration to explore these evolutionary scenarios in gene tree anomaly zones. Our study confirms and expands upon previous work by Zhu and Yang [25] and Long and Kubatko [13] on 3-population trees. By studying migration in a 4-population tree, we can apply our model to more real-world scenarios.

In the future, we hope to generalize our generator matrix algorithm to include more populations and more sequences per population. Furthermore, we will perform simulation studies to generate sequence data in the gene tree anomaly zone in order to compare the performance of difference phylogenetic inference methods. Our work provides a methodology for studying gene-tree-species-tree conflict with more populations. It is important to preform analytical studies of gene-tree-species-tree conflict such as these because it not only informs future simulation studies, but it buttresses the findings of those simulation studies.

References

[1] Elizabeth S. Allman, James H. Degnan, and John A. Rhodes. “Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent”. In:

Journal of Mathematical Biology 62.6 (2011), pp. 833–862. ISSN: 03036812. DOI: 10.1007/s00285-010-0355-7. arXiv: 0912.4472.

[2] David F. Anderson. *Lecture notes on Stochastic Processes with Applications in Biology*. Mar. 2017.

[3] James H. Degnan. "Anomalous unrooted gene trees". In: *Systematic Biology* 62.4 (2013), pp. 574–590. ISSN: 10635157. DOI: 10.1093/sysbio/syt023.

[4] James H. Degnan and Noah A. Rosenberg. "Discordance of species trees with their most likely gene trees". In: *PLoS Genetics* 2.5 (2006), pp. 762–768. ISSN: 15537390. DOI: 10.1371/journal.pgen.0020068.

[5] A.W.F Edwards. "Estimation of the Branch Points of a Branching Diffusion Process". In: *Journal of the Royal Statistical Socity. Series B (Methodological).* 32.2 (1970), pp. 155–174.

[6] Scott V. Edwards et al. "Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics". In: *Molecular Phylogenetics and Evolution* 94 (2016), pp. 447–462. ISSN: 10959513. DOI: 10.1016/j.ympev.2015.10.027. URL: http://dx.doi.org/10.1016/j.ympev.2015.10.027.

[7] Tomáš Flouri et al. "Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent". In: *Molecular biology and evolution* 35.10 (2018), pp. 2585–2593. ISSN: 15371719. DOI: 10.1093/molbev/msy147.

[8] Jody Hey. "Isolation with migration models for more than two populations". In: *Molecular Biology and Evolution* 27.4 (2010), pp. 905–920. ISSN: 07374038. DOI: 10.1093/molbev/msp296.

[9] Jody Hey and Rasmus Nielsen. "Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis". In: *Genetics* 167.2 (2004), pp. 747–760. ISSN: 00166731. DOI: 10.1534/genetics.103.024182.

[10] Nicholas J. Higham. "The Scaling and Squaring Method for the Matrix Exponential Revisited Nicholas J . Higham Manchester Institute for Mathematical Sciences School of Mathematics The University of Manchester". In: *SIAM Review* 51.4 (2009), pp. 747–764.

[11] J. F C Kingman. "The coalescent". In: *Stochastic Processes and their Applications* 13.3 (1982), pp. 235–248. ISSN: 03044149. DOI: 10.1016/0304-4149(82)90011-4.

[12] Liang Liu, Lili Yu, and Scott V. Edwards. "A maximum pseudo-likelihood approach for estimating species trees under the coalescent model". In: *BMC Evolutionary Biology* 10.1 (2010), p. 302. ISSN: 14712148. DOI: 10.1186/1471-2148-10-302. URL: http://www.biomedcentral.com/1471-2148/10/302.

[13] Colby Long and Laura Kubatko. "The effect of gene flow on coalescent-based species-tree inference". In: *Systematic Biology* 67.5 (2018), pp. 770–785. ISSN: 1076836X. DOI: 10.1093/sysbio/syy020. arXiv: 1710.03806.

[14] Wayne P. Maddison. "Gene trees in species trees". In: *Systematic Biology* 46.3 (1997), pp. 523–536. ISSN: 10635157. DOI: 10.1093/sysbio/46.3.523.

[15] Martın Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[16] S. Mirarab et al. "ASTRAL: Genome-scale coalescent-based species tree estimation". In: *Bioinformatics* 30.17 (2014), pp. 541–548. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu462.

[17] Siavash Mirarab and Tandy Warnow. "ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes". In: *Bioinformatics* 31.12 (2015), pp. i44–i52. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv234.

[18] Cleve Moler and Charles Van Loan. "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later". In: *SIAM Review* 45.1 (2003), pp. 3–49. ISSN: 00361445. DOI: 10.1137/S00361445024180.

[19] Richard Nichols. "Gene trees and species trees are not the same". In: *Trends in Ecology & Evolution* 16.7 (2001), pp. 358–364.

[20] Bruce Rannala and Ziheng Yang. "Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci". In: *Genetics* 1656.August (2003), pp. 1645–1656.

[21] Bruce Rannala and Ziheng Yang. "Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference". In: *Journal of Molecular Evolution* 43.3 (1996), pp. 304–311. ISSN: 0022-2844. DOI: 10.1007/pl00006090.

[22] Yuan Tian and Laura S. Kubatko. "Expected pairwise congruence among gene trees under the coalescent model". In: *Molecular Phylogenetics and Evolution* 106 (2017), pp. 144–150. ISSN: 10959513. DOI: 10.1016/j.ympev.2016.09.023. URL: http://dx.doi.org/10.1016/j.ympev.2016.09.023.

[23] J. Wakely. *Coalescent Theory: An Introduction*. Macmillan Learning, 2016. ISBN: 9780974707754. URL: https://books.google.com/books?id=x30RAgAACAAJ.

[24] Chi Zhang et al. "Bayesian inference of species networks from multilocus sequence data". In: *Molecular Biology and Evolution* 35.2 (2018), pp. 504–517. ISSN: 15371719. DOI: 10.1093/molbev/msx307.

[25] Tianqi Zhu and Ziheng Yang. "Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow". In: *Molecular Biology and Evolution* 29.10 (2012), pp. 3131–3142. ISSN: 07374038. DOI: 10.1093/molbev/mss118.

[26] Emile Zuckerkandl and Linus Pauling. "Molecules as documents of history". In: *Journal of Theoretical Biology* 8.2 (1965), pp. 357–366.