ABSTRACT

LEE, HUI-JIE. Advances in Bayesian Inference for Species Divergence Times. (Under the direction of Jeffrey Thorne.)

The amount of sequence divergence depends on the product of the rate of molecular evolution and the time since common ancestry. When only sequence data are available, these rates and times are confounded. Conventionally, rates and times are disentangled by supplementing interspecific sequence information with additional time information (e.g., information from fossils or from "ancient" DNA). Alternative approaches of separating evolutionary rates and divergence times are explored by supplementing interspecific sequence data with information about rates in this dissertation.

The possibility of employing mutation data to assist with the inference of divergence times is investigated in Chapter 2. The basic idea is that the rate of selectively neutral molecular evolution equals the mutation rate and, with the advent of high-throughput sequencing, it is becoming increasingly practical to characterize mutation rates and patterns from parent-offspring data and/or from mutation accumulation lines. The BEAST (Bayesian Evolutionary Analysis Sampling Trees) software package has been modified so that divergence times can be estimated from interspecific sequence data together with information on mutation that is collected by high-throughout sequencing. The approach can be applied when evolutionary rates are treated as being constant through time and also when the evolutionary rates themselves evolve. This research direction will have particular value when studying parts of the tree of life that have poor fossil records and also when studying how mutation rates evolve. Analytical and asymptotic properties of the mutation model are discussed in Chapter 3, and some of the modeling extensions are described in Chapter 5.

Because point mutations can arise from a variety of molecular mechanisms, the change in substitution rate over time can differ among substitution types. The molecular clock assumption is relaxed to different degrees for different substitution types in Chapter 4. To quantify the rate variation among substitution types and to improve divergence time estimation, substitution histories of homologous sequences according to their posterior distribution are sampled. These histories allow 'substitution lengths' for each combination of branch and substitution type on a phylogenetic tree to be inferred. The degree to which substitution types vary in their tendency to change rates can then be examined. In addition, this information can be used to infer divergence times. A nice feature of the approach is that context-dependent substitution (e.g., CpG hotspots) can be accommodated. In accord with previous findings, there is comparatively little rate variation over time for CpG transitions. Separately considering these relatively clock-like substitutions yields improved divergence time estimates.

Advances in Bayesian Inference for Species Divergence Times

by
Hui-Jie Lee

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2014

APPROVED BY:

Sujit Ghosh	Eric Stone
Brian Wiegmann	Jeffrey Thorne Chair of Advisory Committee

DEDICATION

To my family.

BIOGRAPHY

Hui-Jie Lee was born in Kaohsiung, Taiwan. She graduated from National Tsing Hua University in Taiwan with a bachelor degree in Life Science in 2007. In August 2008, Hui-Jie came to North Carolina State University to pursue a doctoral degree in Statistics. She earned an en-route master degree in Statistics in 2010 and started to do research with her advisor, Dr. Thorne in late 2010. She will complete her Ph.D degree in 2014.

ACKNOWLEDGEMENTS

Studying a Ph.D is like wandering through a deep dark tunnel. I feel lost occasionally and do not know where I am heading to and when it is going to end. On the other hand, I feel I am lucky to have the luxury of time to pursue my research. It is hard to imagine that I have spent nearly four years on studying topics that my family and most of my friends do not understand. Until now, my family still do not understand what topic I have been studying, but they always support me whatever I do. I am grateful for their unconditional love and support.

This work would not have been possible without the advice and support of my generous advisor, Jeffrey Thorne. He is the kindly light that leads me through the darkness. There are no words to express my appreciation for him. I would like to thank all members in Thorne group for giving me inputs for my work, in particular Kuangyu Wang and Benjamin Redelings for helping me understand the complicated BEAST software. In addition, I would also like to thank my collaborators Hirohisa Kishino and Nicolas Rodrigue for their constructive comments. Last but not least, I want to thank all my committee members, Sujit Ghosh, Eric Stone, Brian Wiegmann for their advices.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

All life forms are genetically related. Charles Darwin first proposed a common ancestor of all organisms on earth in his *On the Origin of Species*, where he wrote "probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed"[8]. He further suggested the concept of "Tree of Life" such that the pattern of the relationship between species is shaped like a tree. He wrote:

> The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year represent the long succession of extinct species...As buds give rise by growth to fresh buds, and these if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.

To resolve at least part of the "Tree of Life", the study of evolutionary history of organisms, phylogenetics, arises. Understanding the evolutionary process is an essential problem in biological science, and it can be a foundation stone of many advanced studies. These include explaining the similarities and differences among species, testing hypotheses of the generation of biological diversity, understanding the spread of disease and more. The purposes of phylogenetic studies are not only tracing of lineages and history among groups of species, but also estimating the time of divergence between species since they last shared a common ancestor. This dissertation focuses on estimation of species divergence times, and the tree topology is assumed to be known throughout. Note that even though some phylogenetic software (e.g. multidivtime[9, 10], PAML[11, 12]) require users to provide a phylogenetic tree, BEAST[13, 14] and MrBayes[15] can jointly estimate the tree topology and divergence times.

1.1 Phylogenetic Trees

Ever since Charles Darwin illustrated the genealogical relationships among species by a tree, biologists have been using tree diagrams to represent evolution, and such a tree is called a phylogenetic tree, or phylogeny. From a mathematical perspective, a phylogenetic tree is a graph composed of nodes (vertices) and branches (edges), where the nodes represent the taxonomic groups, and the branches correspond to the evolutionary history between descendent and ancestral nodes.

Nodes on a phylogenetic tree can be further classified into three categories: tips, internal nodes, and a unique root. Tips, or terminal nodes, represent extant taxa, and internal nodes represent extinct hypothetical ancestors of descendant nodes. The unique root corresponds to the most recent common ancestor of all the taxa at the tips of the tree. A collection of species (nodes) on a tree can form a clade, or monophyletic group, which means it contains an ancestral species and all the descendants of that ancestor, including extant or extinct descendants. For example, Node A, B and D of Rooted tree 1 in Figure 1.1 form a clade.

A phylogenetic tree can be rooted or unrooted, with or without specifying the position of the root. An unrooted tree only illustrates the relationship between terminal nodes, but it does not make any assumptions about ancestry. Figure 1.1 depicts three rooted trees that share the same unrooted tree topology. The root position is not specified on the unrooted tree, and an unrooted tree can be converted to a rooted one by placing the root on any location on any of its branches.

The shape, or the branching pattern of a tree, is called the topology. A tree topology represents the evolutionary relationships between species so that its tips are labeled, but it does not depict time or evolutionary distance information. In contrast, a labeled history refers to a rooted tree of branching events (topology) in which the speciation events (internal nodes) occur in a temporal order. In other words, nodes are rank-ordered in time in a labeled history. A particular tree topology may have one or more possible labeled histories.

A phylogenetic tree that only shows the topology but contains no information about the branch length is called a cladogram. A phylogenetic tree that represents branch lengths and topology information is called a phylogram. Rooted tree 1 in Figure 1.1 is generated by two speciation events that occurred at times t_0 and t_1 ago. Note that the current day has time 0 and history has time greater than 0 (i.e., $t_0, t_1 > 0$.). The lengths of branches (b_0, b_1, b_2, b_3) correspond to the amount of evolution accumulated in these branches. In molecular phylogenetics, branch lengths are usually measured by the expected number of character changes per sequence site. If the branch lengths are proportional to evolutionary time so that present-day taxa are equidistant from the root, then the tree is called a chronogram or an ultrametric tree. To transform from a phylogram to a chronogram, external information about time and

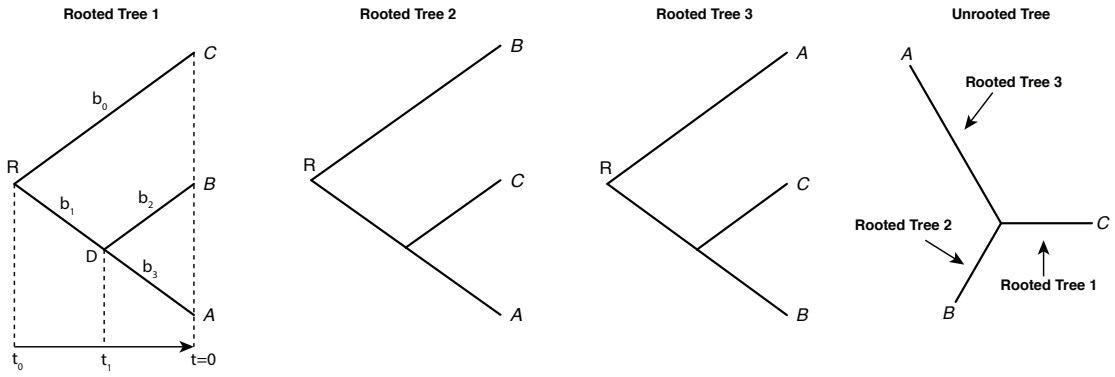


Figure 1.1: The rooted trees (three left columns) that share the same unrooted tree (right column) for 3 species, A, B, and C. The most recent common ancestor of these species, R, is shown in the rooted trees, but not in the unrooted tree. The direction of evolutionary time is represented by an arrow in Rooted Tree 1. Node D corresponds to the most recent common ancestor of Node A and B, and the time to their most recent common ancestor is t_1. b_0, b_1, b_2, b_3 are the lengths of branches, which represent the amount of evolution accumulated in these branches.

assumptions about the rate of evolution are usually needed.

Typically, phylogenetic trees are bifurcating, meaning that every internal node has only two immediate child nodes. In this case, a tree of s extant species will have $(2s-3)$ branch lengths in the unrooted tree and $(s-1)$ times of the internal nodes on the rooted tree to be estimated. For divergence time estimation, the $(s-1)$ times of speciation events are the parameters of most interest.

1.2 The Rate of Evolution and the Molecular Clock Hypothesis

Zuckerkandl and Pauling observed the differences in hemoglobin between different species were roughly proportional to the time to common ancestor[16], and later they proposed the molecular clock hypothesis[17]. The molecular clock hypothesis presumes a clock-like behavior of molecular change. To be precise, it postulates that the rate of molecular evolution is constant over time and shared among evolutionary lineages. Therefore, the amounts of character change from the root to every tip on a phylogenetic tree are expected to be equal with a molecular clock. Figure 1.2 represents phylogenetic trees with and without the molecular clock assumption. The branch lengths of these trees exhibit the evolutionary distances between species, and they are in the unit of the expected number of character changes per site. The phylogenetic tree constructed from Gene 1, which has a single constant rate throughout the tree, has all tips

being equidistant from the root. Since extant species are equidistant with respect to time from their common ancestor, a genetic equidistant tree that resulted from a constant rate over time can be translated into absolute geological times with the help of fossil records. On the other hand, if the rate of evolution varies between branches, then the expected number of substitutions accumulated on each branch would be different. The tree constructed from Gene 2 in Figure 1.2 is an example of rate variation over lineages.

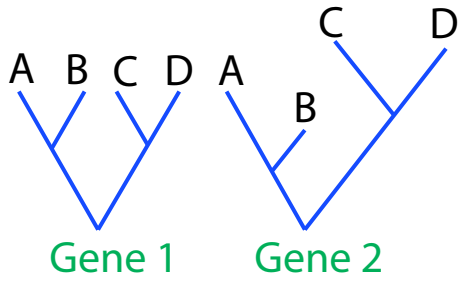


Figure 1.2: Two phylogenetic tress with and without the molecular clock assumption. The phylogenetic tree constructed from Gene 1, which has a single constant rate throughout the tree, has all tips being equidistant from the root. With rate of evolution varying between branches, the tree constructed from Gene 2 has unequal root-to-tip branch lengths.

The rate of evolution is determined by the mutation rate and the probability of fixation (survival) of a mutation. Mutations, which produce differences between species, are the driving forces of species evolution. Mutations can occur owing to different mechanism, including spontaneous cellular processes (e.g. deamination, errors introduced during DNA repair, replication errors, etc.) and induced mutations caused by mutagens (e.g. chemical agents, radiation, etc.). A mutation is said to reach fixation if the frequency of the newly arisen allele is at 100% within the population. A mutation fixation hence leads to the substitution. That is to say, factors that influence the mutation rate and the fixation probability would affect the rate of evolution.

Mutation rates can be affected by factors such as nucleotide sequence context (long repeats, GC content, etc.). For instance, CpG dinucleotides mutate at a higher rate than other types of dinucleotides[18, 19, 20, 21, 22]. CpG dinucleotides are mutational hotspots because DNA methylation occurs frequently at the cytosine in vertebrates. In humans, between 60% and 90% of CpG are methylated on the cytosine ring[23]. Spontaneous deamination at the methylated cytosine causes CpG to TpG (or to CpA in the complementary strand), resulting in a deficiency of CpG and an excess of TpG[24, 25, 26, 27]. The proportion difference of CpG dinucleotides in exons and noncoding DNA (10% and less than 4%, respectively) causes the difference in

mutation rates between these regions[22]. Subramanian and Kuman[22] hypothesized that the differential CpG content could explain the enormous difference in the evolutionary divergence between the fourfold degenerate sites in exons and noncoding (pseudogenes, introns and intergenic regions) DNA.

On the other hand, structural or functional features of the sequence can influence the fixation probability of a mutation. Functional importance leads to low evolutionary rates[28]. For example, the amino acids on the surface region of the alpha and beta hemoglobin chains evolve 10 times faster than those in the functionally important heme group, which carries oxygen molecules[29]. The probability of fixation of a mutation depends on the fitness effect of this mutation as well as the effective population size (N_e). Every genotype has equal probability to reach fixation under no natural selection. Thus, the probability that a new mutation fixes is $\frac{1}{N_e}$ in a haploid population. Indeed, Kimura has proposed the neutral theory that predicts the substitution rate is equal to the mutation rate per site per generation for selectively neutral sites, independent of the population size[30, 31]. In other words, the substitution rate would remain constant throughout the tree provided that the mutation rate stays the same over time in neutrally evolving regions. It therefore leads to the study of neutral molecular clock, which directly provides information to date phylogenies. In contrast, advantageous and deleterious mutations have higher and lower probabilities of fixation than neutral mutations, respectively, because of the selection pressure. Consequently, sequences subject to purifying selection (negative selection) would evolve slower than neutrally evolving sequences, while positively selective sequences would evolve faster.

The rate of evolution can vary among lineages as well. Goodman[32] and Goodman et al.[33] estimated the substitution rates from amino acid sequence data and suggested the rate of molecular evolution decelerated in higher primates (hominoids). They then proposed the hominoid rate-slowdown hypothesis and conjectured that increasing generation times may have been an essential parameter underlying these observations. Furthermore, Laird et al. [34] and Kohne et al. [35] calculated the rates of nucleotide substitution by DNA hybridization and concluded that the rate in rodents is much higher than in primates. The rate difference diminishes after correcting the rate in terms of generation times. These observations lead to the proposal of the generation time effect hypothesis to explain the rate variation among lineages. This hypothesis postulates the molecular clock runs faster in species with a shorter generation time, for they will have more germ-line cell divisions per unit time and thus more errors may occur during DNA replication.

Other explanations have been suggested to account for the rate heterogeneity among lineages in addition to the generation time hypothesis. Britte[36] observed that sea urchins evolve at a fast nucleotide substitution rate as do rodents and drosophila. Unlike rodents and drosophila, sea urchins have longer generation times and thus shorter generation times do not imply fast

rates of evolution. However, because sea urchins produce gigantic numbers of gametes, they also have a large number of germ-line DNA replications per year. Therefore, he suggested that this large number of germ-line DNA replications per year could be a part of the reason of the rapid rates of evolution in these groups of organisms. He further suggested that changes in the effectiveness of DNA repair mechanisms are likely to cause the rate difference among lineages. Apart from the above two hypotheses, Martin and Palumbi proposed the metabolic rate hypothesis[37] that can justify some exceptions the generation time hypothesis fails to explain. In particular, in spite of shorter generation times (about 5 to 10 years), whales show a low rate of nuclear DNA evolution compared to primates[38]. The metabolic rate hypothesis postulates that species with higher metabolic rates will have higher rates of DNA damage through oxygen radicals, and thus are likely to have higher DNA substitution rates[37].

Owing to the complexity of the process of molecular evolution, none of the above three hypotheses could completely account for patterns of rate heterogeneity among lineages. These hypotheses are not mutually exclusive. Furthermore, other biological factors, such as nucleotide composition and recombination rates, can affect the rates of substitutions. For example, some substitutions (especially C to T and G to A) occur at a higher rates than others; G and C tend to change to A and T more frequently than from the opposite direction in pseudogenes[39, 40]. As a result, the molecular clock assumption might be more appropriate for one region in the genome than another. Likewise, it might be relatively suitable for a group of closely related species with similar physiology and life history traits. Recognizing this, maximum likelihood and Bayesian methods that allow deviation from a clock have been developed[41, 9, 42, 43, 10, 44, 45, 46, 47]. These "relaxed clock" models enable evolutionary rates to vary from lineage to lineage. Some relaxed clock models exhibit autocorrelation over time as would be the case if sources of variation in rate are heritable. With autocorrelation, closely related species might evolve at more similar rates than more distantly related species[48]. Different forms of relaxed clock models are discussed in Section 1.4.3.

1.3 Statistical Inference of Phylogenies

1.3.1 Data

Most often, phylogenies are inferred from morphological or molecular sequence data. Molecular sequences (e.g., DNA sequences) consist of genetic information inherited from ancestors, and have the advantage of straightforward quantification owing to the discrete character traits (i.e., 4 nucleobases, 20 amino acids). This dissertation will consider only DNA sequence data. There are 4 types of DNA-bases: A (adenine), C (cytosine), G (guanine) and T (thymine). Nucleotide bases can be classified in two groups based on the structure of the molecules: purines (A, G)

and pyrimidines (T, C). Changes of a purine to another purine (A $\leftrightarrow$ G) or a pyrimidine to another pyrimidine (T $\leftrightarrow$ C) are called transitions. In contrast, transversions refer to changes between a pyrimidine and a purine (T, C $\leftrightarrow$ A, G).

Figure 1.3: An example of a nucleotide sequence alignment from 7 species of primates. Identical nucleotides at a site are shaded if the number of matching nucleotides is higher than 50%.

To reconstruct a phylogenetic tree, a set of well-aligned homologous sequences from the species of interest is needed. Figure 1.3 is an example of a DNA sequence alignment from 7 species of primates, and we assume that these sequences have been derived from a common ancestral sequence. This alignment consists of 27 sites, and we can view it as a 7×27 matrix. For phylogenetic analysis, the data is the homologous sequence alignment of s species with n sites, and it is denoted by an $s \times n$ matrix $\mathbf{X} = \{x_{jk}\}$, where x_{jk} is the k-th site in the j-th sequence. The $-$ symbol in the alignment is a gap (null base), and it is most often treated as a missing value. Usual assumptions made for phylogenetic inference include:

1. The uncertainty of the alignment can be neglected.
2. Each column in the alignment is derived from a single common ancestor.
3. Each site in an alignment evolves independently on an evolutionary tree that is shared with other sites.

Starting from a common ancestral sequence, nucleotide substitutions have occurred and accumulated with time to produce divergent descendant sequences. Sequence similarity or dissimilarity hence can be a measure for sequence divergence. For example, 2 sites out of 27 sites are different between human and orangutan in Figure 1.3, so there is $\frac{2}{27} = 7.4\%$ sequence divergence between human and orangutan. In contrast, there is $\frac{3}{27} = 11.1\%$ sequence divergence between human and rhesus. This suggests that human is more closely related to orangutan than rhesus. Even though the expected proportion of different sites between related sequences increases with time, multiple substitutions or multiple hits at the same site may

happen and may obscure one another. For instance, suppose the ancestral sequence had G at Site 1 and it changed to C to produce Sequence 1, and later it changed from C back to G to produce Sequence 2. In this case, no substitutions could be detected between the ancestral sequence and Sequence 2, even though two substitutions actually occurred. In other words, the number of pairwise sequence differences can underestimate the number of substitutions that occurred since common ancestry, and the expected proportion of different sites is not linear with evolutionary time. Therefore, a probabilistic model of sequence change is required to correct for multiple hits (substitutions) at a site.

1.3.2 Models of DNA sequence change

Models of DNA sequence change provide a mathematical description of the process of DNA sequence change over time. The process of nucleotide substitution is stochastic and is conventionally described by a Markov chain. Since there are 4 possible nucleotide states {T,C,A,G} at each site in a DNA sequence, the Markov chain is characterized by a 4×4 substitution rate matrix $\mathbf{Q} = \{q_{ij}\}$. The element in the substitution rate matrix q_{ij} is the instantaneous rate of change at a site from nucleotide type i to type j. From the theory of finite-state continuous time Markov chains,

$$q_{ii} = -\sum_{j=1, j\neq i}^{4} q_{ij}, \qquad \text{where} \quad q_{ij} \geq 0 \quad \text{for} \quad i \neq j,$$

and thus

$$\sum_{j=1}^{4} q_{ij} = 0, \qquad \forall i.$$

In other words, the row sums of the rate matrix are zero. Note that the instantaneous rate does not depend on time, which means that the substitution process is assumed to be time-homogeneous here. Moreover, because substitutions from any types of nucleotide to any other should be unrestricted, the nucleotide substitution process is modeled as an irreducible Markov chain. There exists a unique stationary distribution of an irreducible Markov chain. That is, the Markov chain ultimately converges to an equilibrium distribution, no matter at what state it started. Denote the stationary distribution of the chain as $\pi = \{\pi_T, \pi_C, \pi_A, \pi_G\}$, where π_i is the proportion of nucleotide type i at equilibrium. A large number of models of DNA sequence change have been introduced by making different assumptions about frequencies of the 4 types of nucleotides at equilibrium and about rates of changing from one type to another.

The simplest model of nucleotide change is the Jukes and Cantor model (JC69 model)[49], which assumes that all substitutions are equally likely with rate λ and all nucleotides occur at

the same frequency ($\pi_i = \frac{1}{4}$ for all i). The substitution rate matrix is

$$\mathbf{Q} = \{q_{ij}\} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix} \end{array}.$$

By distinguishing the rate of transitions and transversions, Kimura proposed an alternative model of nucleotide change (K2P)[50]. Denote the substitution rates of transitions as α and transversions as β. The substitution rate matrix is as follows.

$$\mathbf{Q} = \{q_{ij}\} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix} \end{array}.$$

The above two models assume four nucleotides have equal base composition at equilibrium. The Felsenstein (F84)[51, 52] and Hasegawa-Kishino-Yano (HKY85)[53, 54] models relax this assumption and allow nucleotides to be present in different frequencies. Three additional free parameters are introduced to represent the nucleotide frequencies: π_T, π_C, π_A, and $\pi_G = 1 - \pi_T - \pi_C - \pi_A$. These two models are similar, but their parametrization is different. The substitution rate matrix of the HKY85 model is

$$\mathbf{Q} = \{q_{ij}\} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{bmatrix} -(\alpha\pi_C + \beta\pi_R) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & -(\alpha\pi_T + \beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha\pi_G + \beta\pi_Y) & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & -(\alpha\pi_A + \beta\pi_Y) \end{bmatrix} \end{array},$$

where $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$ are the frequencies of pyrimidines and purines, respectively. All of the models above have a property known as time-reversibility, meaning that the expected number of substitutions from type i to type j is equal to the number of substitutions from type j to i. It is formulated by

$$\pi_i q_{ij} = \pi_j q_{ji}, \qquad \forall i \neq j. \tag{1.1}$$

The GTR (general time-reversible) model is the most general time-reversible model of nu-

cleotide change for a single sequence position. The GTR model described by Simon Tavaré[55] allows unequal base composition and a different rate for each pair of nucleotide substitutions. Instead of only having two parameters for rates of transitions and transversions, the GTR model has $3 \times 2 = 6$ parameters ($\alpha, \beta, \gamma, \delta, \epsilon$ and ε) for rates changing from type i to type j. The substitution rate matrix is

$$\mathbf{Q} = \{q_{ij}\} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{bmatrix} - & \alpha\pi_C & \beta\pi_A & \gamma\pi_G \\ \alpha\pi_T & - & \delta\pi_A & \epsilon\pi_G \\ \beta\pi_T & \delta\pi_C & - & \varepsilon\pi_G \\ \gamma\pi_T & \epsilon\pi_C & \varepsilon\pi_A & - \end{bmatrix} \end{array},$$

where diagonal entries are values that make each row sum to zero.

All of the aforementioned substitution models place constraints on the GTR model and thus they can be viewed as nested within the GTR model. In general, the GTR model with more parameters fits the data better than simple models, but the estimate of branch length has a larger variance as well. For closely related species, the assumptions in the nucleotide substitution models are usually acceptable and therefore simple models produce similar estimates of branch lengths to those under more complex models[56]. However, for distantly related species, branch lengths, especially for longer branches, tend to be underestimated when simple nucleotide substitution models are employed[57, 58]. This represents the bias-variance trade-off in phylogenetic inference. A simple substitution model underfits the data, and it is usually biased but has low variance. A more complicated substitution model is less biased but loses the precision of estimates. Therefore, the solution is not to use a complicated model but to select a simple model that still fits the data sufficiently well.

Based on different theoretical foundations, there are various criteria for model selection. These include: the likelihood ratio test (LRT), the Akaike information criterion (AIC)[59], the Bayesian information criteria (BIC)[60] and the Bayes factor (BF). jModelTest 2 is one of the most popular software packages for comparing substitution models with some of the above criteria[61]. A thorough discussion of these criteria is beyond the scope of this dissertation. In fact, research is still ongoing around estimating the Bayes factor due to the difficulty in computing the marginal likelihoods, which usually do not have closed-form expression (e.g. Fan et al.[62], Xie et al.[63], Baele et al.[64]).

Given the substitution rate matrix $\mathbf{Q}$ of a Markov chain, the transition probability matrix $\mathbf{P}(t)$ at time t can be derived by solving the following differential equation

$$\mathbf{Q} = \left.\frac{d\mathbf{P}(t)}{dt}\right|_{t=0}, \tag{1.2}$$

with the condition that $\mathbf{P}(0) = \mathbf{I}$. The solution is

$$\mathbf{P}(t) = \{p_{ij}(t)\} = e^{\mathbf{Q}t}, \tag{1.3}$$

where $e^{\mathbf{Q}t}$ is the matrix exponential of $\mathbf{Q}t$, and it can be defined by means of a power series:

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \mathbf{I} + \sum_{k=1}^{\infty} \frac{(\mathbf{Q})^k t^k}{k!}. \tag{1.4}$$

The element in the transition probability matrix $p_{ij}(t)$ is the probability of having nucleotide type j at time t given nucleotide type i at time 0. For example, the transition probability matrix for the JC69 model is

$$\mathbf{P}(t) = \{p_{ij}(t)\} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix} \end{array},$$

where $p_0(t) = \frac{1}{4} + \frac{3}{4}\exp(-4\lambda t)$ and $p_1(t) = \frac{1}{4} - \frac{1}{4}\exp(-4\lambda t)$. There is no closed-form solution of the transition probability matrix for the GTR model, so it is usually solved numerically. The transition probability matrix $\mathbf{P}(t)$ can be used to calculate the likelihood of an alignment on a tree, which will be discussed in Section 1.3.4.

Note that $\mathbf{Q}$ and t occur in the form of a product in Equation 1.3. Therefore, it is conventional to scale the rate matrix $\mathbf{Q}$ by a scaling factor η so that the average rate of substitution is 1, where

$$\eta = -\frac{1}{\sum_{i=1}^{4} \pi_i q_{ii}}.$$

For instance, the scaling factor would be $\frac{1}{3\lambda}$ in the JC69[49] model. After scaling, the rate matrix $\mathbf{Q}' = \eta\mathbf{Q}$ would represent the relative substitution rates only. Hence, the time (or equivalently, branch length) t is measured by the expected number of substitutions per site. For simplicity, $\mathbf{Q}$ will denote either the rate matrix without scaling if the time t is measured by the chronological time, or the rate matrix after scaling if the time (branch length) t is measured by the expected number of substitutions per site.

Using the Markov property, the Chapman-Kolmogorov equation can be derived. It is

$$p_{ij}(t+s) = \sum_{k} p_{ik}(t) p_{kj}(s). \tag{1.5}$$

That is, the probability of changing from nucleotide type i to type j in time $(t+s)$ is the sum over all possible transition states k of the probability of changing from nucleotide type i to k in time t and then from k to j in time s. In other words, by modeling the substitution process as Markovian, all possible paths leading from nucleotide type i to j in time $(t+s)$ are considered and thus multiple hits are automatically corrected.

Using the transition probability instead, Equation 1.1 implies

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \qquad \forall t. \tag{1.6}$$

For an ancestral and a descendant sequence, the time-reversibility property means that it is not possible to distinguish which sequence is ancestral from the two sequences alone. This is because the substitution process looks exactly the same whether time goes forward or backward. In fact, by combining Equation 1.5 and Equation 1.6, the probability of observing type i in Sequence 1 and type j in Sequence 2 that are diverged from a most recent common ancestor t and s time units ago can be simplified as follows.

$$\sum_{k=0}^{\infty} \pi_k p_{ki}(t) p_{kj}(s) = \sum_{k=0}^{\infty} \pi_i p_{ik}(t) p_{kj}(s) = \pi_i p_{ij}(t+s). \tag{1.7}$$

For the purpose of calculating likelihoods, this implies that one of the sequences can be viewed as ancestral and the other can be viewed as the descendant that resulted after $(t+s)$ units of evolution.

1.3.3 Models of Among-Site Rate Variation

The rate of nucleotide substitution can vary among sites and among loci. Divergence times can be underestimated if rate variability is ignored, because multiple substitutions that have occurred at mutation hotspots may not be detected[65]. As a matter of fact, failing to accommodate the rate variation among sites has greater effect on divergence time estimation than failing to properly model the relative rates of nucleotide substitution process[66].

Variable substitution rates among sites can be modeled by either discrete or continuous distributions.

Discrete-distribution model Hasegawa et al.[67] first proposed to divide sites into two categories, *invariant sites* with zero rate of change, and variable sites with a constant rate. Yang[68] later generalized the two-rate model into the discrete-rates model with K rate classes. In this model, the rate for a site is assumed to come from one of K rate classes. A class has rate value r_k and probability p_k, where $k = 1, 2, ..., K$. In practice, K is typically a small integer less than 10. Two more constraints are made in order to reduce the dimensionality of the parameter

space. First, p_k is usually set to $\frac{1}{K}$ for all k. Second, the mean rate is 1: $\sum_k p_k r_k = 1$, which makes the branch length be the expected number of substitutions per site averaged over all sites. Since the average rate is still 1, the rate r_k represents how fast class k evolves compared to the average rate across all sites. The resulting substitution rate matrix at a site with rate r_k is simply $r_k\mathbf{Q}$, where $\mathbf{Q}$ is the original substitution rate matrix shared among all sites. The scaling of rate r_k to the substitution rate matrix is the same for the continuous-distribution model described below.

Continuous-distribution model Another approach is to have the rate for any site be randomly sampled from a more biologically plausible continuous distribution. Uzzell and Corbin[69] first modeled rates among sites by a gamma distribution, and the gamma model was further developed by many researchers, e.g. Nei and Gojobori[70], Jin and Nei[71], Tamura and Nei[72]. A suffix '+Γ' is added to substitution models if rate variation among sites is modeled by a gamma distribution, such as HKY+Γ. In the gamma-distributed rates model, the rate at each site is assumed to be independently and identically distributed according to the gamma distribution

$$f(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1}, \qquad \alpha > 0, \quad \beta > 0, \quad r > 0,$$

with mean $E(r) = \frac{\alpha}{\beta}$ and variance $Var(r) = \frac{\alpha}{\beta^2}$. To avoid the use of too many parameters, the mean of the distribution is restricted to 1 by setting $\alpha = \beta$, and the variance is hence $\frac{1}{\alpha}$. The extent of rate variation among sites can thus be measured by α. Large values of α ($\alpha > 1$) indicate most sites have rates around 1 and there is minor rate variation, whereas small values of α ($\alpha \leq 1$) mean a high degree of rate variation among sites. When α is approaching infinity, the rates become identical for all sites.

To reduce the computation for the gamma rate model, Yang[66] approximated the gamma distribution with K equal probability rate classes, so that the mean or median in each class is used to represent all rates in that rate class. This model is called the *discrete-gamma model*. A suffix '+Γ_k' is added to substitution models if the discrete-gamma model is used to account for rate heterogeneity among site, where the subscript k is the number of rate classes. In this model, the probability of falling into each class p_k is $\frac{1}{K}$, and the rate value for each class r_k is determined by the gamma distribution. Yang[66] has shown that as few as four rate classes can greatly improve model fit to real data, but model fit often does not improve much by having many more rate classes than four.

1.3.4 Likelihood Calculation of a Tree

Related sequences can be viewed as a realization of random variables with probability distribution generated from a model of evolution. The likelihood of the data (alignment) can be

calculated given the tree topology, branch lengths, and a model of sequence change. Let the parameters in the model of sequence change be θ, and the branch lengths be $\mathbf{T}$. θ and $\mathbf{T}$ are the parameters in the model, and the maximum likelihood estimates of θ and $\mathbf{T}$ can be found by optimizing the likelihood function given the tree topology. The maximum likelihood estimates of the phylogeny can be found by repeating the same procedure for each possible tree topology and searching for the topology that maximizes the likelihood of the data. Two assumptions are often made for computing the likelihood of an alignment on a tree:

1. Each site evolves independently.

2. Given the state of the site at the beginning of the branch, evolution is independent in each branch of the tree.

By making the first assumption, the likelihood of an alignment on a tree is the product of the likelihood of data at individual sites. In other words, the likelihood calculation can be simplified by first computing likelihoods at individual sites separately and then taking the product of individual likelihoods. The second assumption basically says that the process of nucleotide substitution is Markovian.

Using the same notation in Section 1.3.1, the alignment is denoted by an $s \times n$ matrix $\mathbf{X} = \{x_{jk}\} = \{\mathbf{x}_k\}$, where x_{jk} is the k-th site in the j-th sequence, and $\mathbf{x}_k$ is the k-th column in the alignment. To illustrate likelihood calculation for a single site on the tree of Figure 1.4, we will temporarily suppress the notation concerning sequence position and will instead write x_i for the state of node i. This means $x_i \in \{A, T, C, G\}$, where $i = 0, 1, ..., s, ..., 2s-2$. Let t_i be the branch length of a branch ending with node i, and let $\mathbf{T} = \{t_i, i = 1, ..., 2s-2\}$. Then the likelihood of the tree for this single site is the sum, over all possible states of the internal nodes:

$$\begin{aligned} &Pr(x_1 = A, x_2 = C, x_3 = G, x_4 = C, x_5 = C|\mathbf{T}, \theta) \\ &= \sum_{x_0}\sum_{x_6}\sum_{x_7}\sum_{x_8} Pr(x_1 = A|x_6, t_1)Pr(x_2 = C|x_6, t_2)Pr(x_3 = G|x_8, t_3)Pr(x_4 = C|x_7, t_4) \\ &\qquad Pr(x_5 = C|x_7, t_5)Pr(x_6|x_0, t_6)Pr(x_7|x_8, t_7)Pr(x_8|x_0, t_8)Pr(x_0). \end{aligned} \tag{1.8}$$

The transition probability $Pr(x_i|x_j, t_i) = p_{ji}(t_i)$ and the equilibrium probability $\pi = Pr(x_0)$ are specified by the model of sequence change ($\mathbf{P}(t)$ and π) in Section 1.3.2. The amount of computation of the likelihood in Equation 1.8 can be economized by using Felsenstein's dynamic programming based "pruning" algorithm[73, 74]. The idea is to identify common factors and move the summation signs in Equation 1.8 to the right as far as possible. In this

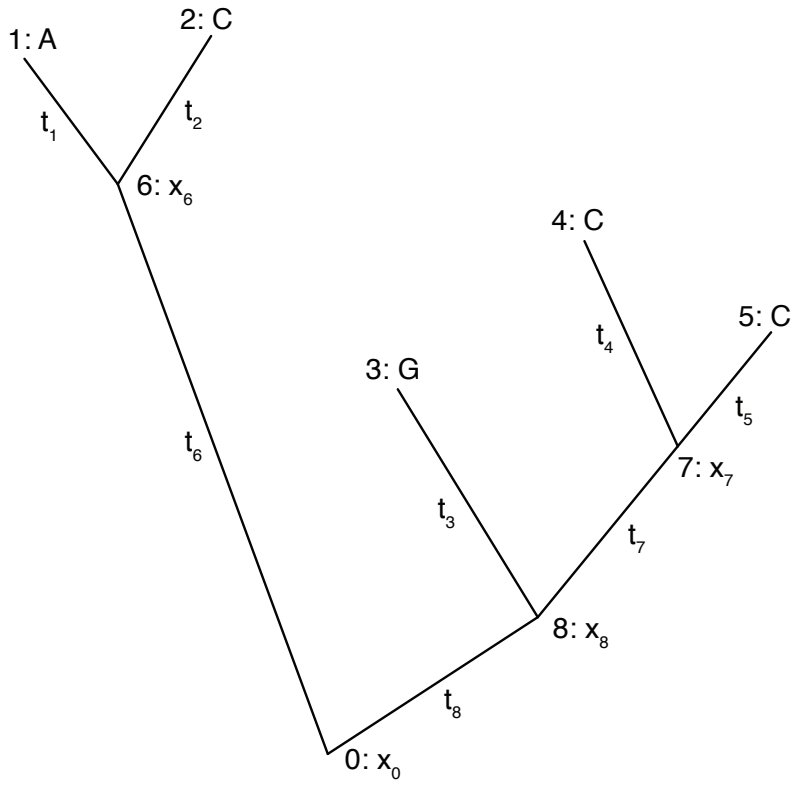


Figure 1.4: The tree used in the discussion of the likelihood calculation. Nodes are labeled with numbers and their states (nucleotide types). The t's are the branch lengths.

way, Equation 1.8 becomes

$$
\begin{aligned}
&Pr(x_1 = A, x_2 = C, x_3 = G, x_4 = C, x_5 = C|\mathbf{T}, \theta) \\
&= \sum_{x_0} \left(\sum_{x_6} Pr(x_6|x_0, t_6) Pr(x_1 = A|x_6, t_1) Pr(x_2 = C|x_6, t_2) \right) \times \\
&\left\{ \sum_{x_8} Pr(x_8|x_0, t_8) Pr(x_3 = G|x_8, t_3) \left(\sum_{x_7} Pr(x_7|x_8, t_7) Pr(x_4 = C|x_7, t_4) Pr(x_5 = C|x_7, t_5) \right) \right\}.
\end{aligned}
\tag{1.9}
$$

Equation 1.9 suggests that the summation over all possible states of ancestral nodes are taken after summing over all possibles states for its descendent nodes. In other words, the pruning algorithm performs a postorder tree traversal for computing the likelihood of a tree. This algorithm converts a calculation that grows exponentially with the number of internal nodes to one that grows linearly with the number of nodes. Define $L_i(x_i)$ as the probability of observing

data at the tips that are descendants of node i given that node i has state x_i. Then $L_6(C)$ is the probability of observing $x_1 = A$ and $x_2 = C$ given that node 6 has state C. Provided that node i is a tip, this definition implies that $L_i(x_i) = 1$ if its observed state is actually x_i and $L_i(x_i) = 0$ otherwise. For an internal node i with descendant nodes j and k,

$$L_i(x_i) = \left(\sum_{x_j} Pr(x_j|x_i, t_j) L_j(x_j) \right) \times \left(\sum_{x_k} Pr(x_k|x_i, t_k) L_k(x_k) \right). \tag{1.10}$$

The overall likelihood for this single site can then be calculated from the tips of the tree to the root, and it is given by

$$Pr(\mathbf{x}_k|\mathbf{T}, \theta) = \sum_{x_0} Pr(x_0) L_0(x_0). \tag{1.11}$$

Finally, the likelihood of the alignment on a tree can be computed by the product of individual likelihoods.

$$L = \prod_{k=1}^{n} Pr(\mathbf{x}_k|\mathbf{T}, \theta). \tag{1.12}$$

If rate heterogeneity among sites is modeled by a discrete distribution, then the likelihood is the weighted average over K rate classes

$$L = \prod_{k=1}^{n} \left[\sum_{h=1}^{h=K} p_h \times Pr(\mathbf{x}_k|\mathbf{T}, \theta, r = r_h) \right], \tag{1.13}$$

where r_h is the rate value for class h and p_h is the probability of falling in class h. $Pr(\mathbf{x}_k|\mathbf{T}, \theta, r = r_h)$ can be computed by using the transition probability matrix $\mathbf{P}(t) = e^{r_h \mathbf{Q} t}$ such that the substitution rate matrix $\mathbf{Q}$ is multiplied by r_h. Similarly, if rate heterogeneity among sites is modeled by a continuous distribution, the likelihood is the average over the rate distribution

$$L = \prod_{k=1}^{n} \left[\int_0^{\infty} f(r) Pr(\mathbf{x}_k|\mathbf{T}, \theta, r = r) dr \right], \tag{1.14}$$

where $f(r)$ is the probability density of the rate (e.g. a gamma probability density) and $Pr(\mathbf{x}_k|\mathbf{T}, \theta, r = r)$ can be calculated from the transition probability matrix $\mathbf{P}(t) = e^{r \mathbf{Q} t}$ so that the rate matrix $\mathbf{Q}$ is scaled by r. An algorithm for calculation of the above likelihood of a tree under a continuous-rates model has been provided by Yang[75].

1.3.5 Identifiability Problems in Divergence Time Estimation

The transition probability matrix $\mathbf{P}(t)$ used to calculate the likelihood of an alignment has the form in Equation 1.3. Note that the rate matrix $\mathbf{Q}$ and time t occur in the form of a product, and thus only the effect of the product can be seen.

For example, consider two sequences separated by time t under the JC69 model. The total rate of substitution of any type of nucleotide is 3λ as specified in the rate matrix. Therefore, the expected number of substitutions per site, or the branch length, is $3\lambda t$. The probability that the two sequences are different at a site at time t is $p = 3p_1(t) = 1 - p_0(t)$, where $p_1(t) = \frac{1}{4} - \frac{1}{4}\exp(-4\lambda t)$ and $p_0(t) = \frac{1}{4} + \frac{3}{4}\exp(-4\lambda t)$ as defined in Section 1.3.2. Suppose there are x out of n sites that are different between the two sequences, then the estimated probability of different sites is $\hat{p} = \frac{x}{n}$. By equating the expected and the estimated probability of different sites, we have

$$\hat{p} = \frac{3}{4} - \frac{3}{4}\exp(-4\hat{\lambda}\hat{t}), \tag{1.15}$$

which suggests that only the product of λ and t, or the branch length $3\lambda t$, can be estimated from the data. In other words, it is not possible to estimate rates and times separately without information other than sequence data because there are problems with identifiability.

Let $\{f(\mathbf{X}|\theta) : \theta \in \Theta\}$ be a family of probability distributions for a model completely specified by parameters θ. Parameters θ of the model are *identifiable* if distinct values of θ correspond to distinct probability distributions. That is, if there exists some values $\theta_1 \neq \theta_2$ such that they give the same value of the likelihood function $f(\mathbf{X}|\theta_1) = f(\mathbf{X}|\theta_2)$ for all data $\mathbf{X}$, then the parameters θ of the model are *non-identifiable* or *unidentifiable*. In the previous example of divergence time estimation between two sequences under JC69 model, a combination of low rate $\lambda = 0.1$ and longer time $t = 5$ will give the same probability of different sites p as a combination of high rate $\lambda = 0.5$ and shorter time $t = 1$.

Additional information about rates and times are needed in order to disentangle rates and times in the sequence data. One possibility is to use the molecular clock assumption of a constant rate to date divergence times. Under the clock assumption, if a particular divergence time can be calibrated either by a fossil record or biogeographical evidence or sample dates for a quickly evolving virus, then all other times on the tree can be estimated. However, substitution rate is affected by the population size and the life history of a species (e.g. generation time and body size). For these and other reasons, the clock assumption may be violated[37]. Moreover, uncertainties in the age of the fossil or uncertainties in assignment of a fossil on the phylogeny remain a problem and need to be considered carefully. Therefore, a general framework capable of not only incorporating rate variation among lineages and fossil uncertainties, but also handling non-identifiability problems in the model, is in demand.

Unlike frequentist inference which introduces additional constraints on the model to make it

identifiable, "unidentifiability causes no real difficulty in the Bayesian approach" as suggested by Lindley[76]. As long as a "proper" prior is specified on all the parameters, Bayesian inference can handle unidentified complicated models[77]. The Bayesian method for phylogenetics has several extra advantages as suggested by Huelsenbeck et al.[78]. First, it is based on the likelihood function, so it shares some strengths of maximum likelihood methods, e.g. explicit model assumptions. Second, prior knowledge about rates and times and uncertainties about fossil records can be easily and naturally accommodated under the Bayesian framework. Third, the posterior distributions can be approximated by MCMC and they have easy interpretations, especially for summarizing tree topologies and explaining credible intervals.

However, if an improper prior is used because no additional information about rates or times is available, it may lead to an improper posterior[77]. In other words, the Bayesian approach provides a convenient way to incorporate prior knowledge from rates and times, but it cannot make something out of nothing. It is because identifiability is not a problem of a particular estimation procedure, but a property of the model. Therefore, good fossil records, carefully chosen substitution models, models of rate evolution and priors are important for estimating divergence times. In the next section, Bayesian inference of divergence times will be described in detail.

1.4 Bayesian Inference of Divergence Times

The development of Bayesian MCMC methods in phylogenetics was originated by three groups of authors, Yang and Rannala[79], Li et al.[80, 81], and Mau and Newton[82, 83]. From then on, the incorporation of more complex and realistic evolutionary models, the relaxation of molecular clock assumption, the accommodation of uncertainties in fossil records, tree topologies, and other parameters have been made due to the feasibility of Bayesian approach.

1.4.1 General Bayesian Framework

The problem of divergence time estimation can be formulated in the general Bayesian framework as follows. Let $\mathbf{X}$ be the aligned molecular sequences of s species with n sites. Let Θ include all parameters in the substitution models (e.g. transition/transversion ratio, shape parameter for discrete-gamma rates model). Let τ be the tree topology for s species, which can be either specified before the analysis or estimated with divergence times. Let $\mathbf{t}$ be a vector of the $(s-1)$ divergence times and $\mathbf{r}$ a vector of rates. The rates can be either for the $(2s-2)$ branches or for the $(2s-1)$ nodes, or even for the $(2s-2)$ branches and the root, depending on the implementation. Denote Φ as the parameters of the model of rate change among lineages, and denote Ω as the parameters of the prior on times and tree topology. Then the joint posterior

distribution is given as

$$f(\mathbf{t}, \mathbf{r}, \tau, \Theta, \Phi, \Omega|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{t}, \mathbf{r}, \tau, \Theta) f(\mathbf{r}|\mathbf{t}, \tau, \Phi) f(\mathbf{t}, \tau|\Omega) f(\Theta, \Phi, \Omega)}{f(\mathbf{X})}. \tag{1.16}$$

The term $f(\mathbf{X}|\mathbf{t}, \mathbf{r}, \tau, \Theta)$ is the likelihood of the alignment $\mathbf{X}$ given the tree topology τ, which can be calculated under a nucleotide substitution model and a rate heterogeneity model, as discussed in Section 1.3.4. Note that the branch length is the product of branch rate and time duration of that branch. Therefore, approximation may be needed in order to obtain the average rate for the branch. The term $f(\mathbf{r}|\mathbf{t}, \tau, \Phi)$ is the prior probability of rates given the tree, which is specified by a model of rate change among lineages. Note that the prior on rates is conditioned on times because the evolutionary rate is described as a function of time, $r(t)$. That is, the rate of evolution can change along the branches of the tree. The prior on times and tree topology $f(\mathbf{t}, \tau|\Omega)$ can be a birth-death prior that specifies both tree topology and times or a prior on times with discrete uniform prior on tree topology. Besides, if fossil calibration is available, this information is incorporated in the prior for times. And finally, $f(\Theta, \Phi, \Omega)$ is the joint prior for the parameter Θ, Φ, Ω.

Priors reflect knowledge or assumptions about values of parameters, but the information for specifying priors is often not available, especially for nuisance parameters (e.g. parameters in the substitution models). Some common priors include: beta or Dirichlet distribution for proportions, gamma or lognormal distribution for parameters with positive support, discrete uniform prior for tree topologies, and more. Hyperparameters of prior distributions can have their own prior distribution to express uncertainty in hyperparameters. Uninformative flat priors for nuisance parameters are usually used in phylogenetic analysis. Sometimes, this can lead to improper posteriors. Therefore, prior sensitivity analysis should be conducted to test the effect of prior distributions. Priors on rates and times will be discussed next.

1.4.2 Priors on Times and Tree Topologies

Inference of divergence times can be done on a fixed tree or with inference of the tree topology as well. Therefore, priors on divergence times can be handled in two different ways: specifying priors on times with a fixed tree topology, or jointly specifying priors on times (or branch lengths) and tree topologies.

Fixed Tree Topology A fixed tree topology can be either inferred from another analysis or specified by knowledge about the evolutionary relationship between the species. Different types of priors on times can be specified conditioned on a fixed tree topology.

1. Uniform Prior

The simplest prior for times is the uniform prior. Suppose $t_{D(i)}$ and $t_{A(i)}$ are the divergence times of the oldest descendant node and the ancestral node of node i, respectively. Then the prior on time of node i, t_i, is uniformly distributed on the interval $[t_{D(i)}, t_{A(i)}]$. Fossil information indicates that a species arose in some time interval, and thus fossil calibrations for node times are sometimes incorporated by placing additional uniform priors on certain nodes. Such fossil calibrations are called "hard" bounds. Yang and Rannala[84] showed that hard bounds can lead to misleading narrow credible intervals for divergence time estimates when there are conflicts among fossil calibrations or between fossils and sequences.

2. Dirichlet Prior

Kishino et al.[10] intended to provide a statistically simple and flexible prior on times, so they put a gamma prior on the root time with a Dirichlet prior on the proportional times of all other internal nodes relative to root time. Suppose a particular path starting from the root includes k branches ($k-1$ speciation events excluding the root), and let p_i be the proportion of time of branch i relative to the total time of the path. Then the joint prior of the proportions of times in k branches could be a Dirichlet distribution with equal means.

$$f(p_1, p_2, ..., p_k|\alpha) = \frac{\Gamma(k\alpha)}{(\Gamma(\alpha))^k} \prod_{i=1}^{k} p_i^{\alpha-1}, \qquad 0 < p_i < 1, \quad \sum_{i=1}^{k} p_i = 1, \quad \alpha > 0. \tag{1.17}$$

Note that the priors on times proposed by Kishino et al.[10] are specified recursively and the probability density $f(\mathbf{t})$ lacks a closed-form solution, which makes it difficult to implement flexible priors for fossil calibrations.

3. Birth-death Prior

Rannala and Yang[79, 85] used a birth-death process which explicitly models lineage speciation, extinction and taxon-sampling to specify the priors on times. A birth-death process has parameters for birth rate λ, death rate μ, and sampling fraction ρ. Parameter λ and μ are the constant speciation and extinction rates per lineage, respectively. Therefore, the probability of a speciation or an extinction event occurs on a lineage in an infinitesimal time interval Δt is $\lambda\Delta t$ or $\mu\Delta t$, respectively. Parameter ρ is the proportion of extant species included in the phylogeny. Let $\mathbf{t}_{-1} = \{t_2, t_3, ..., t_{s-1}\}$ be the node times excluding the root time t_1. The joint prior of the $(s-2)$ node times $\mathbf{t}_{-1}$ conditioned on the root

time t_1 for a given tree topology τ with s species is

$$f(\mathbf{t}_{-1}|t_1, \tau, \lambda, \mu, \rho) = (s-2)! \prod_{i=2}^{s-1} \frac{\lambda p_1(t_i)}{\nu_{t_1}}, \tag{1.18}$$

where

$$p_1(t) = \frac{1}{\rho} P(0,t)^2 e^{(\mu-\lambda)t}, \tag{1.19}$$

$$\nu_{t_1} = 1 - \frac{1}{\rho} P(0,t_1) e^{(\mu-\lambda)t_1}, \tag{1.20}$$

$$P(0,t) = \frac{\rho(\lambda-\mu)}{\rho\lambda + (\lambda(1-\rho)-\mu)e^{(\mu-\lambda)t}}. \tag{1.21}$$

A Yule process is a special case of birth-death process with zero death rate ($\mu = 0$) and with complete taxon-sampling ($\rho = 1$).

The use of a birth-death prior makes implementation of flexible priors for fossil calibration possible because the prior density can be calculated analytically. Yang and Rannala[84] devised an approach to implement "soft bounds" or flexible distributions to describe uncertainties in fossil information by using the birth-death process. Let $\mathbf{t}_C$ be node times with fossil calibration and $\mathbf{t}_{-C}$ be the node times without fossil information so that $\mathbf{t}_{-1} = \{\mathbf{t}_C, \mathbf{t}_{-C}\}$. Suppose the uncertainties in fossil information are accommodated by a probability density $f(\mathbf{t}_C|C)$, then the joint prior for times from the birth-death process with fossil calibration is

$$f(\mathbf{t}|C) = f(t_1, \mathbf{t}_{-1}|C) = f(t_1) f_{BD}(\mathbf{t}_{-C}|\mathbf{t}_C) f(\mathbf{t}_C|C) = \frac{f(t_1) f_{BD}(\mathbf{t}_{-1}|t_1)}{f_{BD}(\mathbf{t}_C|t_1)} f(\mathbf{t}_C|C), \tag{1.22}$$

where $f(t_1)$ is the prior for the root time, $f_{BD}(\mathbf{t}_{-1}|t_1)$ is from Equation 1.18, and $f_{BD}(\mathbf{t}_C|t_1)$ is the marginal density of calibrated node times from the birth-death prior. The formula for $f_{BD}(\mathbf{t}_C|t_1)$ can be found in Yang and Rannala's paper[84].

Inferred Tree Topology Priors on trees are often set to be uniform over a set of trees. There are $(2s-5)!! = \frac{(2s-5)!}{2^{s-3}(s-3)!}$ possible tip-labeled unrooted trees and $(2s-3)!! = \frac{(2s-3)!}{2^{s-2}(s-2)!}$ rooted trees for s species. On the other hand, there are $\frac{(s-1)!s!}{2^{s-1}}$ labeled histories for s species[86]. Because some topologies may have more than one possible labeled history, there are more labeled histories than rooted tree topologies in general. Uniform priors on tree topologies do not place equal probability on each node ordering. And in fact, Pickett and Randle[87] demonstrated that these priors imply non-uniform priors for clades such that small clades and large clades have higher prior probabilities than mid-sized clades.

1. Flat Priors on Tree Topologies

A flat prior, or discrete uniform prior, places equal prior probability on all possible tip-labeled tree topologies. Then the prior for tree topologies τ is $f(\tau) = \frac{1}{B(s)}$, where $B(s)$ is the number of possible tip-labeled rooted or unrooted trees. Priors on times can then be specified separately by either flat priors or any arbitrary distributions. Huelsenbeck and Ronquist wrote the widely used software MrBayes[15] that places a flat prior on tree topologies with either uniform or exponential priors on branch lengths.

2. Birth-death Prior

Since birth-death priors explicitly model a tree branching process, they are natural as priors for tree topologies. Edwards showed that a Yule process (and thus birth-death process) makes each distinct labeled history equally probable[88]. As a result, the prior for a particular rooted tree (labeled history) τ is $f(\tau) = \frac{1}{B(s)}$, where $B(s)$ is the number of possible labeled histories for s species. The joint prior of a particular labeled history τ and node times $\mathbf{t}$ is then

$$\begin{aligned} f(\mathbf{t}, \tau | t_1, \lambda, \mu, \rho) &= f(\mathbf{t}_{-1} | t_1, \tau, \lambda, \mu, \rho) f(t_1) f(\tau) \\ &= f(\mathbf{t}_{-1} | t_1, \tau, \lambda, \mu, \rho) f(t_1) \times \frac{2^{s-1}}{(s-1)! s!}, \end{aligned} \tag{1.23}$$

where $f(t_1)$ is the prior for the root time and $f(\mathbf{t}_{-1} | t_1, \tau, \lambda, \mu, \rho)$ is from Equation 1.18.

1.4.3 Models of Rate Change Among Lineages and Priors on Rates

By virtue of the development of Bayesian approach in phylogenetics, various models of rate variation among lineages, or so-called the "relaxed clock" models, have been proposed in the Bayesian framework. The models of rate change among lineages can be basically classified into four types: global molecular clock, local molecular clocks, autocorrelated rate models and uncorrelated or independent rate models. Note that these models may model the rates of nodes or branches depending on the implementation details.

Global Molecular Clock

The global molecular clock, or the strict clock, assumes the rate of evolution is constant over time, and therefore there is a single rate r on the tree. A gamma prior is usually used for the single rate.

Local Molecular Clocks

Closely related species usually have similar rate of evolution. By observing a rate slow-down in hominoid evolution, Koop et al.[89] suggested to apply a molecular clock to a

localized set of branch points, which led to the idea of the local clock. Under the local clock, branches are grouped into rate classes so that branches in the same rate class share the same constant rate. The events of rate changes occur infrequently. The local clock model was first implemented by Kishino and Hasegawa[90, 91] in a maximum likelihood framework.

Drummond and Suchard[92] developed a Bayesian random local clock model such that every branch can either keep the same rate of its parent branch or has a rate change to form a new local clock. Let r_i be the rate of branch i on a tree with s species, $i = 1, 2, ..., 2s-2$. And let $\delta = \{\delta_1, ..., \delta_{2s-2}\}$ be a vector of indicators variables such that $\delta_i = 0$ if the rate of branch i is the same as its parent rate and $\delta_i = 1$ otherwise. Then the rate of branch i is

$$r_i = c(\boldsymbol{\rho}) \times \rho_i = c(\boldsymbol{\rho}) \times \rho_{A(i)} \times \phi_i = r_{A(i)} \times \phi_i, \tag{1.24}$$

where $A(i)$ refers to the parent branch of branch i, ϕ_i is a branch-specific rate multiplier, and $c(\cdot)$ is a normalization constraint. Priors on rates are specified through the sum of the indicator variables δ and the rate multipliers $\boldsymbol{\phi} = \{\phi_1, ..., \phi_{2s-2}\}$. Let K be the sum of indicator variables, $K = \sum_{i=1}^{2s-2} \delta_i$. The prior on K is a truncated poisson distribution with a hyperparameter λ, which is the expected number of rate changes on the tree. And the prior on rate multiplier ϕ_i is a gamma distribution with mean 1 and variance $\psi\phi_i$, where ψ is a hyperparameter that quantities the variance. When $\delta_i = 0$ so that the rate of branch i is the same as its parent branch, the rate multiplier has expected value 1 and zero variance. In contrast, when $\delta_i = 1$ so that the rate of branch i changes, then the rate multiplier has mean 1 and variance ψ.

Autocorrelated Rate Models

In order to approximate rate drifts over time in a continuous fashion but still account for the fact that closely related species are expected to have similar rates, autocorrelated rate models have been developed. A geometric Brownian motion model which recursively models rate changes among lineages is proposed by Thorne et al.[9] and Kishino et al.[10] and later modified by Rannala and Yang[47]. In the implementation of Kishino et al.[10], a gamma prior is selected for the rate at the root. Then the rate of a node given the rate of its ancestral node has a lognormal distribution with the constraint that descendant rate and ancestral rate have equal mean. Let r_i and $r_A(i)$ be the rate of a node and its ancestral node, respectively. And let Δt_i be the time duration separating the two nodes.

Then the prior on r_i conditioned on $r_{A(i)}$ is given by

$$f(r_i|r_{A(i)}, \Delta t_i) = \frac{1}{r\sqrt{2\pi\Delta t_i \nu}} \exp\left\{-\frac{1}{2\Delta t_i \nu}\left(\log\left(\frac{r_i}{r_{A(i)}}\right) + \frac{\Delta t_i \nu}{2}\right)^2\right\}, 0 < r_i < \infty, \nu \geq 0, \tag{1.25}$$

where ν is the variance parameter quantifying the amount of rate autocorrelation. A large ν means the rate variation among lineages is large, while a small ν implies strong rate autocorrelation among lineages. Equivalently, the lognormal autocorrelated rate model implies the log rate of a node, $\log(r_i)$, given the log rate of its ancestral node, $\log(r_{A(i)})$, follows a normal distribution.

$$\log(r_i)|\log(r_{A(i)}) \sim N\left(\log(r_{A(i)}) - \frac{\Delta t_i \nu}{2}, \Delta t_i \nu\right). \tag{1.26}$$

Consequently, this model is also called a geometric Brownian motion rate model, since the logarithm of the rate changes with time in terms of a Brownian motion. Let $\mathbf{r} = \{r_1, r_2, ..., r_{2s-1}\}$ be the set of node rates on a tree with s species, and let r_1 be the root rate with prior $f(r_1)$. Let $\Delta\mathbf{t} = \{\Delta t_1, ..., \Delta t_{2s-1}\}$ be the time durations between node i and its parent node. Then the joint prior for node rates is

$$f(\mathbf{r}|\Delta\mathbf{t}) = f(r_1) \prod_{i=2}^{2s-1} f(r_i|r_A(i), \Delta t_i), \tag{1.27}$$

where $f(r_i|r_A(i), \Delta t_i)$ is from Equation 1.25, and $\Delta\mathbf{t}$ can be calculated given the divergence times $\mathbf{t}$ on a tree τ. An additional prior for the variance parameter ν needs to be specified as well.

Uncorrelated or Independent Rate Models

An alternative approach to allow rates to drift over time is to impose no rate autocorrelation on adjacent branches of the tree. Rannala and Yang[47] implemented the *independent rates model* so that branch rates are drawn independently from an identical distribution. Let $f(r_i)$ be the prior of branch rate r_i, which has the same distribution for every branch. Then the joint prior of branch rates $\mathbf{r} = \{r_1, ..., r_{2s-2}\}$ on a tree of s species is

$$f(\mathbf{r}) = \prod_{i}^{2s-2} f(r_i). \tag{1.28}$$

Instead of directly sampling rates from an exponential or a lognormal distribution, Drummond et al.[46] discretized one of the distributions into $(2s-2)$ rate categories and assigned rate categories to branches by sampling without replacement from the discretized

distribution. This model is called the *uncorrelated rate model*. The rate for category c is calculated by

$$r_c = F^{-1}\left(\frac{c-0.5}{2s-2}\right), \tag{1.29}$$

where $F^{-1}(\cdot)$ is the inverse distribution function of lognormal or exponential. Lepage et al.[93] note that the uncorrelated rate model is defined on the average rate on each branch, so rates within each branch can still be autocorrelated. And in fact, it is difficult to derive the stochastic process of rate drift analytically so that the resulting average rates on branches of the tree are distributed as an exponential or lognormal distribution.

1.5 BEAST: Bayesian Evolutionary Analysis Sampling Trees

In this section, BEAST (Bayesian evolutionary analysis sampling trees)[13, 14] is described because the methods for incorporating mutation data, which will be introduced in next chapter, is implemented in BEAST. Among many Bayesian Markov chain Monte Carlo (MCMC) software in phylogenetics, BEAST is one of the most commonly-used programs. It is not only because various models are implemented so that divergence times estimation, reconstruction of demographic history, phylogeographic analysis, and other related molecular evolutionary analyses can be done together, but also it has a user-friendly graphical interface.

The key feature of BEAST is that it assumes either the strict or relaxed clock and explicitly models the rate of evolution on each branch, which allows rooted trees with timescales to be inferred. Moreover, in addition to nucleotide and amino acid sequences, it can analyze phenotypic traits and accommodate flexible distributions for fossil calibrations.

BEAST contains a large collection of source files written in the Java programming language. It is object-oriented and its classes are organized hierarchically into packages so that adding new features to BEAST can be accomplished by adding new classes which utilize existing methods. In other words, codes can be reused and it is no need to reinvent the wheel when implementing new models.

BEAST takes an XML file, which records the data and model settings, as input and produces two output files containing the posterior samples of parameter values and trees. To implement a model of mutation data as in the next chapter, in addition to a class for model implementation, a corresponding parser class is necessary in order to read instructions from XML inputs. The ideal situation is to put all newly developed classes into plugin objects and allow users to install the plugin without recompiling BEAST. The goal of BEAST 2[94], a completely rewrite of BEAST, is to make everything in BEAST a plugin so that it is easy to extend. However, the autocorrelated lognormal rates model is removed in BEAST 2 and thus implementation of our mutation model is only done in BEAST.

For the sake of generating XML inputs, summarizing and visualizing the outputs, a suite of programs are distributed along with BEAST. They are BEASTGen, BEAUti, LogCombiner, TreeAnnotator, Tracer[95] and FigTree. BEAST and its relatives are open source and available at http://beast-mcmc.googlecode.com/.

The procedure for conducting an analysis in BEAST is as follows. First, prepare a sequence alignment in NEXUS format (a widely used data format for phylogenetic analysis). FASTA format (a widely used format in bioinformatics for storing DNA and/or protein sequence data) or BEAST XML format can be transformed to NEXUS format by BEASTGen. Second, create an input XML file by BEAUti. Additional editing may be required in order to incorporate more complex models. Third, run the analysis in BEAST to produce two output files, a "log" file for sampled parameter values and a "trees" file for sampled trees. Output files from multiple identical runs can be combined by LogCombiner. Resampling posterior states at lower frequency can be done by LogCombiner as well. Tracer and TreeAnnotator can remove burn-in and summarize information from posterior samples in "log" and "trees" files, respectively. Simultaneously, MCMC diagnosis can be performed by Tracer. And finally, the resulting tree from TreeAnnotator can be visualized by FigTree.

Chapter 2

Estimating species divergence times with molecular sequence data but without fossils

2.1 Introduction

Mutations are the substrate for evolution. They generate genetic variation that evolutionary forces, including natural selection and genetic drift, can act upon. The degree of genetic similarity between species reflects the evolutionary relationships between them. However, the amount of evolution accumulated between two species, which is conventionally expressed by the expected number of sequence substitutions per site between their homologous sequences, depends on both the rate of evolution and the time since divergence. Interspecific comparison of molecular sequences can provide information about amount of evolution, but evolutionary rates and times since common ancestry are confounded. Therefore, to be able to date species divergence times, a method that can disentangle times from rates is required.

Under the molecular clock hypothesis[17], which assumes that all lineages have the same rate of sequence change, DNA sequence data along with fossil or geological information allow evolutionary rates to be inferred. The inferred evolutionary rates can then illuminate divergence dates on portions of the phylogeny where direct fossil evidence is unavailable. Without the assumption of a constant rate among lineages, rates and times can be separated to some extent via prior knowledge about their values[9, 10, 42, 43, 44, 46, 92]. Much research has been focused on better ways of supplementing molecular sequence data with information about times in divergence times estimation. Progress has been achieved with regard to better treatment of fossil evidence[84, 47] as well as with the analysis of serially sampled sequence data[96, 97, 98, 99].

Another direction for improvement is to add independent data about evolutionary rates

to analyses of interspecific sequence data. Owing to the development of "next-generation" sequencing technologies[100], direct measurements of the number of mutations within a known number of generations are gradually becoming available. These include, but are not limited to, pedigree-based whole-genome analysis of human mutation rates[101, 102] and estimation of mutation rate by mutation accumulation lines of *Drosophila melanogaster*[103] and *Caenorhabdits elegans*[104, 105]. While the point mutation rate is not necessarily identical to the nucleotide substitution rate, these rates are tightly linked. Better characterization of point mutation will lead to improved ability to separately estimate nucleotide substitution rates and times since common ancestry. To date, high-throughout sequencing studies have mainly been employed to study mutation in humans and in model genetic systems. In the near future, one can envision mutation rates being estimated for any taxa of interest. This means that mutation data can be collected to help separate evolutionary rates and times even for groups with poor paleontological records.

Nevertheless, the information about mutation rate from extant organisms should be treated with caution. First, rates of mutations are typically measured per generation. To convert the mutation rate per generation to mutation rate per year, the generation time of the species should be estimated. However, mean generation time in the wild can be difficult to estimate in the laboratory. Second, for a variety of reasons, mutation rates vary along the genome (e.g. Benzer[106], Wolfe et al.[107], Matassi et al.[108] and Nachman et al.[21]) and among individuals in the same species (e.g. Conrad et al.[102], Haag-Liautard[103] and Kong et al.[109]). Third, owing to sequencing errors, base-calling algorithms for identifying *de novo* mutations may be subject to false positive and false negative rates.

On top of these issues, many studies have suggested a discrepancy between long-term phylogenetic rates (or substitution rates) and short-term mutation rates (e.g. Ho et al.[110]). One possibility is that mutation rates measured over a short timescale will include slightly deleterious mutations which are likely to be removed by purifying selection over time. Moreover, rate of mutation (and therefore substitution) may change along the evolutionary lineages. The rate variation along lineages may be associated with species characteristics, that include generation times, body size, population size, metabolic rates, etc[37, 111]. Therefore, an attempt to utilize mutation rate for estimating divergence times should consider rate variation along evolutionary lineages as well as whether only neutrally evolving regions are to be analyzed or whether to incorporate natural selection as well as mutation.

According to the Neutral Theory of Motoo Kimura[30, 31], the nucleotide substitution rate is equal to the neutral mutation rate. Hence, mutation data are directly informative about substitution rate under the neutral evolution assumption. Some studies have used the approach that Drummond et al.[46] suggested: The mutation rate information can be incorporated in the analysis by either fixing the rate of evolution or setting the mean of the rate prior distribution

equal to the mutation rate[46]. Focusing on divergence times of closely related species and analyzing only synonymous sites of orthologous loci, Cutter estimated the neutral divergences between pairs of sequences and applied contemporary mutation rates to recover divergence times in nematodes[112]. Obbard et al. used an experimentally inferred neutral mutation rate to provide a informative prior for the substitution rates of the third positions of 4-fold degenerate codons of fruit flies[3]. They set the mean of the uncorrelated lognormal rate prior as the estimated mutation rate of one taxon (*Drosophila Melanogaster*) and they associated a small variance to this lognormal prior for rate.

Inspired by these studies, we suggest a Bayesian "relaxed clock" approach for combining mutation data with interspecific sequence data. We treat mutation information as "data" and let it have its own likelihood. The approach provides a natural way to accommodate uncertainty associated with mutation data when estimating divergence times and substitution rates. In addition, our treatment of mutation data allows incorporation of mutation information from multiple species by simply adding mutation likelihoods for each species. As high-throughout sequence data gets less expensive to collect, we envision that parent-offspring mutation data and mutation accumulation data will become increasingly accessible for non-model systems and increasingly important for disentangling substitution rates and times. Extensions of our approach offer the potential for divergence time estimation when neutral evolution is not assumed and when there is instead a mutation-selection balance.

2.2 Materials and Methods

2.2.1 The Bayesian Framework

Let $\mathbf{X}$ be the aligned molecular sequences of s species, which are related via a bifurcating rooted tree topology τ. Denote $\mathbf{r}$ to be a vector of substitution rates (a scalar under the molecular clock). The vector $\mathbf{r}$ can be either node rates or branch rates depending on the implementation. But for simplicity, let us now assume $\mathbf{r}$ be a collection of branch rates, and $\mathbf{r} = \{r_1, r_2, ..., r_{2s-2}\}$, where $\{r_1, ..., r_s\}$ corresponds to the terminal branches. So we can denote $\mathbf{r} = \{\mathbf{r}_T, \mathbf{r}_{-T}\}$, where $\mathbf{r}_T$ is the set of terminal branch rates, and $\mathbf{r}_{-T}$ is the set of internal branch rates. Let $\mathbf{t}$ be a vector of internal times in the species tree, $\mathbf{t} = \{t_1, t_2, ..., t_{s-1}\}$. We represent the parameters in the substitution model (such as the transition/transversion ratio κ in the HKY model[54]) and hyperparameters in the prior distribution of rates (e.g. the autocorrelation parameters in the lognormal autocorrelated clock) as Ω. Whereas we denote the parameters in the tree prior (such as the birth rate in birth-death process prior) with Θ.

In addition to the aligned sequence data $\mathbf{X}$, an independent set of data that provides information about the mutation rates, referred to as the mutation data $\mathbf{M}$, is incorporated into our

Bayesian analysis. A detailed description of our simple model of mutation data will be given later.

For divergence time estimation, we will construct the joint posterior distribution of the above parameters and tree topology given the molecular sequence data $\mathbf{X}$ and the mutation data $\mathbf{M}$. According to Bayes theorem, we have

$$f(\mathbf{t}, \mathbf{r}, \tau, \Omega, \Theta | \mathbf{X}, \mathbf{M}) = \frac{f(\mathbf{t}, \mathbf{r}, \tau, \Omega, \Theta, \mathbf{X}, \mathbf{M})}{f(\mathbf{X}, \mathbf{M})} \tag{2.1}$$

$$= \frac{f(\mathbf{X}|\mathbf{t}, \mathbf{r}, \mathbf{M}, \tau, \Omega, \Theta) f(\mathbf{M}|\mathbf{t}, \mathbf{r}, \tau, \Omega, \Theta) f(\tau, \Omega, \Theta, \mathbf{t}, \mathbf{r})}{f(\mathbf{X}, \mathbf{M})} \tag{2.2}$$

$$= \frac{f(\mathbf{X}|\mathbf{t}, \mathbf{r}, \tau, \Omega) f(\mathbf{M}|\mathbf{r}) f(\tau, \Omega, \Theta, \mathbf{t}, \mathbf{r})}{f(\mathbf{X}, \mathbf{M})} \tag{2.3}$$

Equation (2.3) stems from Equation (2.2) because we assume the molecular sequence data $\mathbf{X}$ are independent of the mutation data $\mathbf{M}$ and only the rates $\mathbf{r}$ are relevant to the calculation of the likelihood of mutation data $\mathbf{M}$. In Equation (2.3), $f(\mathbf{X}|\mathbf{t}, \mathbf{r}, \tau, \Omega)$ is the likelihood of the aligned sequences given the tree topology and the branch lengths that are specified by the rates and times. With conventional models of sequence change (e.g. the Jukes and Cantor model[49], the HKY model[54]), this likelihood can be computed via Felsenstein's pruning algorithm[74]. The likelihood of the mutation data $f(\mathbf{M}|\mathbf{r})$ will be described below and $f(\tau, \Omega, \Theta, \mathbf{t}, \mathbf{r})$ represents prior information.

Including mutation data in the Bayesian framework attaches the likelihood of the mutation data $f(\mathbf{M}|\mathbf{r})$ to the numerator and changes the denominator from $f(\mathbf{X})$ to $f(\mathbf{X}, \mathbf{M})$ compared to the standard posterior distribution of the parameters given only the sequence alignment $\mathbf{X}$. Because the posterior distribution can be approximated by the Metropolis-Hastings algorithm[113, 114] and because this approximation does not require evaluation of the denominator, adding mutation data only influences the acceptance probability for proposing new rates.

2.2.2 Model of Mutation Data

Suppose there are c out of s species that have mutation data $\mathbf{M}$ available. Then we can further partition the terminal branch rates $\mathbf{r}_T = \{\mathbf{r}_C, \mathbf{r}_{-C}\}$, where $\mathbf{r}_C$ have mutation rate information and $\mathbf{r}_{-C}$ do not. The mutation data for each species consist of a direct measurement of the number of mutations from a known number of generations. We can denote the mutation data as a collection of mutation data for each species $\mathbf{M} = \{(y_1, m_1, g_1), ..., (y_c, m_c, g_c)\}$. The mutation data for species i is (y_i, m_i, g_i), where y_i is the number of mutations from a total of m_i bases scanned, and g_i is the mean generation time for species i.

Let us assume the mutation data from different species are independent and calculation of the likelihood of the mutation data only depends on the corresponding terminal branch rate.

Then the likelihood of the mutation data $\mathbf{M}$ is simply the product of the mutation data from each species.

$$f(\mathbf{M}|\mathbf{r}_C) = \prod_{i=1}^{c} f(y_i|m_i, g_i, r_i). \tag{2.4}$$

There are many levels of complication possible for treating the mutation data. We consider the simplest case here by modeling mutation with a Poisson distribution. The Poisson treatment of the mutation data is consistent with the Jukes and Cantor (JC69) model[49] because it assumes the total substitution rate that affects a sequence does not vary among sequences with shared length. If the mutation data consist only the number of mutations and not of the nucleotide types before and after the mutation events, a Poisson treatment would also be consistent with the Kimura 2-parameter (K2P) model[50] because its total rate of substitution for a site does not vary among the four nucleotide types that could occupy the site. However, for more complicated substitution models, Poisson treatment is simply an approximation.

If the mutation rate at a neutrally evolving site was drawn from a gamma distribution, then the number of substitutions at a site can be fit to a negative binomial distribution[69]. Therefore, one possible extension of the mutation model is to incorporate a negative binomial model for the number of mutations with the gamma distributed rates among sites. Although we are not considering the information here, a more complete treatment of mutation data would consider the information that it can provide about nucleotide frequencies, ratios of transition to transversion rate, and other parameters of the substitution process. Our Poisson treatment only uses mutation data to help separate evolutionary rates and times. A better treatment of mutation data would be particularly worthwhile if the neutral evolution assumption were relaxed so that the mutation process combined with natural selection to shape nucleotide substitution. Substitution models that explicitly treat mutation-selection balance are actively being studied[115, 116, 117, 118, 119, 120] and would very much benefit from careful treatments of mutation data.

We model the observed number of point mutations as a realization of a Poisson distribution with mean being the product of the number of bases scanned and the point mutation rate per site per generation. The mean generation time for each species is assumed known so that we can convert the mutation rate per position per generation to the mutation rate per position per year. Second and as mentioned above, we assume the rate of selectively neutral molecular evolution equals the point mutation rate. Last but not least, we assume the number of mutations that occurred equals the number that were detected. In other words, we do not explicitly consider possible false positives and false negatives. A better treatment of the mutation data would model the uncertainty of base calling.

With this approach, the probability of y_i mutations of species i from a total of m_i bases

assayed is

$$f(y_i|m_i, g_i, r_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \tag{2.5}$$

where λ_i is the mutation rate per generation per sequence scanned for species i. Assuming that mutations are being detected by comparing genotypes of parents and offsprings, we can express λ_i as the product of total bases scanned m_i, the mean generation time in years g_i and the mutation rate per site per year r_i for species i. If changes are detected via mutation accumulation studies, λ_i would include the number of generations represented by the study as an additional factor.

2.2.3 Data Sets

Three empirical data sets were analyzed by applying mutation information. The first set consisted of molecular sequences in intergenic regions from six primate species (Figure 2.1). We selected intergenic regions with the hope that these regions are putatively neutral and hence mutation rates are the same as substitution rates. To demonstrate the utility of calibrating the molecular clock for taxa with rare or poor fossil records, two additional data sets were analyzed. One of these consisted of a subset of previously analyzed coding regions of 6 *Caenorhabditis* species from Cutter[112] (Figure 2.2(a)). The other included 50 nuclear genes from 12 species of the *Drosophila* family (Figure 2.2(b)). This third data set was analyzed by Obbard et al.[3]. They used an uncorrelated lognormal relaxed clock and set their substitution rate prior by using mutation information from *C. elegans*. Neutrally evolving regions are required for equating the mutation rate with the substitution rate. With this in mind, only fourfold degenerate sites were used for the *Drosophila* and *Caenorhabditis* analyses. Although fourfold degenerate sites are not necessarily neutral (e.g. Clemente and Vogl[121], Lawrie et al.[122]), we reasoned that they might experience less natural selection than other sites.

Sequence Data

Primate Data Set Sequences orthologous to four noncoding intergenic segments of human chromosome 21 (coordinates chr21:20,219,730-20,220,080; chr21:27,701,500-27,702,600; chr21:30,278,770-30,279,170; chr21:41,310,000-41,311,700 from the February 2009 human reference sequence GRCh37/hg19 assembly[123]) across 6 species of primates were analyzed (see Figure 2.1 for the phylogeny). By using human as the reference species, all 46-way Multiz[124] records that have no overlap with RepeatMasker[125] were obtained from the UCSC genome browser[123]. Sequences from 6 species of primates were extracted and converted from MAF (multiple alignment format) to FASTA format via the Galaxy server[126, 127, 128]. The final alignment contained 2839 sites.

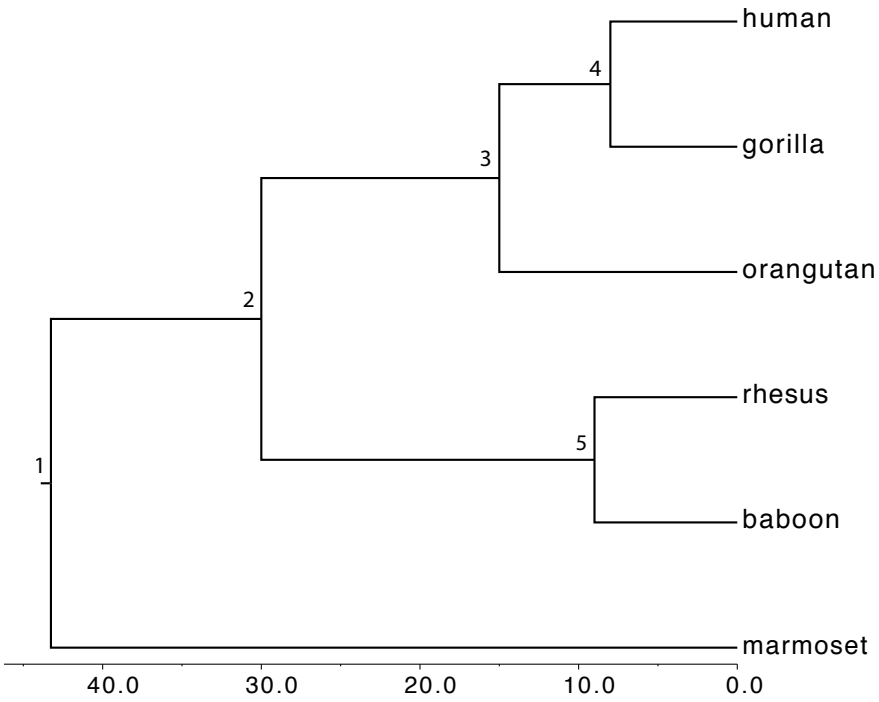
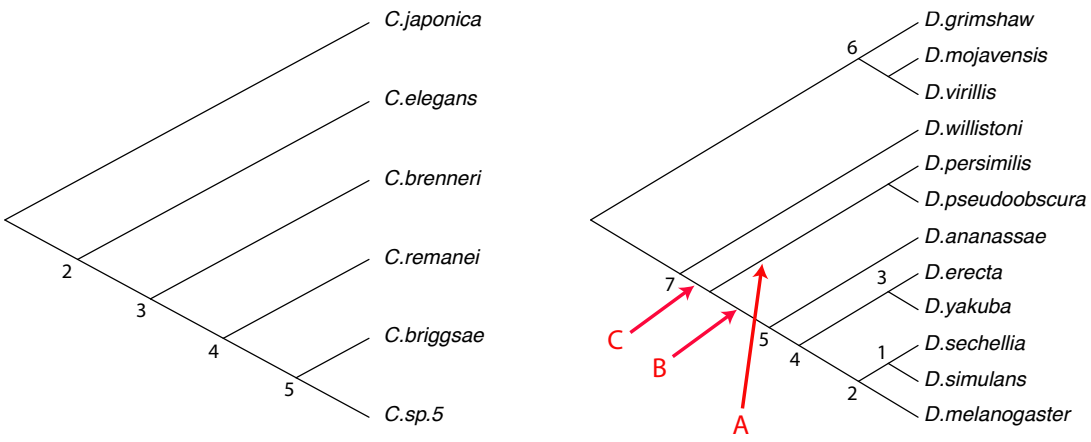


Figure 2.1: Phylogenetic relationships for the 6 primates. One time unit is one million years.

Nematode Data Set The data set from Cutter[112] consisted of 63 orthologous groups that were found in all 6 *Caenorhabditis* species (Figure 2.2(a)) and 7642 orthologous groups for 3 or more taxa. The 'frequency of optimal codons' (Fop) for each locus were kindly provided by Asher Cutter. Any group with high overall codon bias (average Fop over species exceeded 0.4)[129] and with representatives in less than 4 species were filtered out, and thus 21 orthologous groups were left in the analysis. Since Cutter had incomplete genomes at time of his study, an updated and more complete data set was obtained by retrieving curated homologs of *Caenorhabditis elegans* on the WormBase (WS238)[130]. When homologs were not available in both WormBase and Cutter's original data set, the sequences aligned with *Caenorhabditis elegans* by the multiz 6-way whole genome nematode alignment[124] on the UCSC Genome Browser[123] were retrieved. The 21 orthologous sequences with 23406 sites were aligned by Translator X[131] with MAFFT[132], and the fourfold degenerate sites were identified by DnaSP v.5.10.1[133], resulting in 2115 sites in the sequence alignment (Supplementary material: nematode_data.xlsx).

***Drosophila* Data Set** The *Drosophila* data set consisted of 50 protein-coding loci of 12 sequenced species of *Drosophila* (*Drosophila* 12 Genomes Consortium 2007)[7]. These loci were previously analyzed by Obbard et al.[3] and were selected because of their low codon usage



(a) *Caenorhabditis* (b) *Drosophila*

Figure 2.2: Cladograms for 6 species of *Caenorhabditis* and 12 species of *Drosophila* derived from Cho et al.[6] and from Clark et al.[7], respectively. The labeled branches of the *Drosophila* cladograms represent different root placements. Our analysis with the lognormal autocorrelated relaxed clock model placed the root on Branch A if mutation data were used and on Branch B if not. Obbard et al.[3] placed the root on Branch C with the lognormal uncorrelated relaxed clock model and a rate calibration on the mean of the lognormal distribution.

bias. Sequences were downloaded from `ftp://flybase.net/genomes/12_species_analysis/clark_eisen/alignments/` and the four-fold degenerate sites were retrieved by by DnaSP v.5.10.1[133]. The alignment contained 6538 sites. The cladogram of the 12 species is shown in Figure 2.2(b)[134]. Using the program BASEML in PAML[12], maximum likelihood estimates of the branch lengths under the HKY[54] and the discrete-gamma model with four categories[66] for the unrooted tree topology was obtained (Figure 2.3). Note that the branches separating *D. sechellia* – *D. simulans* and *D. persimilis* -*D. pseudoobscura* are very short.

Mutation Data

Primate Data Set The Human mutation data in the analysis were obtained from Conrad et al.[102], which included two separate studies. The CEU study had 45 observed *de novo* mutations from a total of 2,555 Mb diploid bases interrogated. The YRI study had 35 observed *de novo* mutations from a total of 2,549 Mb diploid bases interrogated. The mutation rates after adjusting for false negatives were respectively 1.17×10^{-8} and 0.97×10^{-8} per base per

generation for the CEU and YRI studies. Therefore, the estimated number of point mutations by combining the two studies are calculated as follows.

$$1.17 \times 10^{-8} \times 2,555 \times 2 \times 10^6 + 0.97 \times 10^{-8} \times 2,549 \times 2 \times 10^6$$
$$= 109.24 \text{ estimated mutations in 10,208Mb scanned per generation.}$$

The factor of 2 above converts the diploid base pairs to haploid base pair. In sum, the mutation data in this analysis consisted of 80 detected point mutations and 29 estimated false negatives, added up to 109 point mutations in a total 10,208 millions haploid bases assayed after adjustment of false negatives.

Recent direct genealogical and indirect demographic studies[135, 136] suggest that the average human generation time is around 28 years with female generation time (between 25 and 27 years) being substantially shorter than male (around 31 years). The suggested human generation time is much longer than values often used in population genetic studies. Generation times depend on the ages of first birth and the mortality-weighted age of last birth in a population. Wich et al.[137] indicated that the longer human generation is the result of the late age of first reproduction and the overall low mortality rates. In fact, Matsumura and Forster[136] reported wild chimpanzee generation times from two populations, and the population with a higher mortality rate among adults had a much shorter generation time (19 years) than the other (24 years). We assumed the longer modern human generation time as a result of a recent change in life history (e.g. mortality rate). Therefore, we conservatively assumed human generation time of 20 years similar to some previous studies (e.g. Fu[138] and Nachman et al.[21]). The divergence times estimated with a human generation time of 30 years are reported for reference.

Nematode Data Set The mutation data set for *C. elegans* was from Denver et al.[104]. It consisted of 391 *de novo* point mutations in a total of 584 million assayed nucleotide sites from 10 mutation accumulation (MA) lines and suggested an average base-substitution mutation rate of 2.7×10^{-9} per site per generation. Since these 10 MA lines were collected with varying number of generations and number of sites, we need to convert the mutation data into the format in Section 2.2.2. The mutation rate per nucleotide site per generation μ in this study was calculated with the equation $\mu = m/(nT)$, where m is the number of detected mutations, n is the number of sites, and T is the average number of generations[104]. Therefore, it is equivalent to 391 point mutations detected in $\frac{391}{2.7\times10^{-9}} = 144,814,814,815$ bases scanned in one generation, and this number was used in our analysis.

C. elegans has a minimal generation time about 3 to 4 days in laboratory conditions but the generation time may be much longer in the wild and is uncertain[139]. Following Cutter[112], multiple average generation times (14, 30 and 60 days) were used to convert the mutation rate

per site per generation to the mutation rate per site per million years.

***Drosophila* Data Set** The mutation data set for *D. melanogaster* was from Keightley et al.[140]. It consisted of 174 point mutations detected in 3 MA lines with 72,078,570 sites in an average of 262 generations, which gave an estimate of the mean mutation rate per site per generation as 3.07×10^{-9}. After correcting for the mutations missed by a mutation-calling algorithm, they reported the corrected mutation rate per haploid per site per generation as 3.46×10^{-9}. It is equal to $3.46 \times 10^{-9} \times 3 \times 262 \times 72,078,570 = 196$ estimated mutations in a total of $72,078,570 \times 3 \times 262 = 56,653,756,020$ bases assayed for one generation. Similar to Obbard et al.[3], we assumed 10 generations per year for *D. melanogaster*.

2.3 Results and Discussion

All analyses were performed by the BEAST software[14] that we modified to incorporate our Poisson treatment of mutation data. These Markov Chain Monte Carlo (MCMC) analyses were run for between 10^8 and 2×10^9 iterations depending on the length needed to achieve convergence. The first 10% of the posterior samples were regarded as burn-in and were removed from the analyses. MCMC analyses were each run at least twice. Convergence was assessed by determining that the effective sample size exceeded 200 according to the Tracer software[95].

2.3.1 Divergence Times of Primates

We used the HKY substitution model[54] and the discrete-gamma model of between-site rate variation with 5 rate categories[66]. A lognormal prior with mean 1.0 and standard deviation 1.25 was assigned to the transition/transversion rate ratio κ for the HKY model[54]. An exponential prior with mean 0.5 was assigned to the gamma shape parameter α of the discrete-gamma rate model[66]. Since the divergence time of human and marmoset was suggested 43 Myr ago in TimeTree[4, 5], the prior of the root time was a gamma distribution with mean 43 Myr and standard deviation 6 Myr. The actual marginal priors on divergence times were formed by the combination of the root time prior and a birth-death process (Table 2.1). The mean growth rate and the relative death rate in the birth-death process had the default priors in BEAST[14]. The mean growth rate had a uniform $(0, 100000)$ prior and the relative death rate had a uniform $(0, 1)$ prior. An autocorrelated lognormal relaxed clock was employed to analyze the primate data. We estimated the substitution rate with a strict clock and a calibration point of 43 Myr on the root time in the BASEML program in PAML package[12]. It allowed us to assign an exponential prior with mean (and also standard deviation) 0.0014 per site per Myr for the rate of the root. As suggested in Kishino et al.[10], the mean and the standard deviation

of the gamma prior for the autocorrelation parameter ν were both set $1/43 = 0.024$ because the prior mean for the root age was 43 Myr.

The posterior means and 95% CIs of the divergence times, the mean rate and the rate of the human branch are shown in Table 2.1. First, the human mutation data affected the priors on the rates but had no effect on the priors on the times. Second, the inferred divergence times for all the nodes were substantially older with the human mutation data than without.

These earlier divergence dates were consistent with the results from Langergraber et al.[141]. In their study, instead of directly estimating divergence times with a model of mutation data, they post-processed previously published divergence time estimates with new yearly mutation rates. They converted the experimentally-derived human mutation rate per site per generation to a human mutation rate per site per year. They did this by fixing the human generation time to be 29 years. Since mutation rate estimates were unavailable for chimpanzee and gorillas, Langergraber et al.[141] assumed the mutation rate per site per generation was identical between humans, gorillas and chimpanzees. By further assuming generation times of 19 years for gorillas and 25 years for chimpanzees, Langergraber et al. could derive mutation rates per site per year for gorilla and chimpanzee. Similarly, they derived a mutation rate range for the most recent common ancestor of human and chimpanzee by assuming a generation time of between 25 and 29 years for this ancestor. Finally, they re-calibrated the previously published divergence times between human, gorilla and chimpanzee (5.95 Myr from Scally et al.[142] and 6.69 Myr from Dutheil et al.[143]) by the new mutation rate to conclude a range of the divergence time of $(8.31, 17.79)$ or $(9.35, 20.00)$. Their estimated divergence times were earlier than previously estimates[142]. Simultaneously, Gibbons reported the new mutation rate would put the split of human and orangutan at between 34 million years ago and 46 million years ago[144]. Clearly there is a gap between the date of fossils and the date estimated from the molecular sequence data, and the changes of life history (e.g. body size, generation time) may only be parts of the answer.

2.3.2 Divergence Times of Nematodes

The data set was analyzed under both the strict clock and the autocorrelated lognormal relaxed clock. The HKY substitution model[54] was adopted as was discrete-gamma among-site rate variation model with 4 categories[66]. Nematode fossils are rare and only a very few studies have attempted to date divergence times between nematodes. By analyzing proteins from 12 and 15 eukaryote genomes, including mouses, primates, fruit flies, and more, Blair et al.[145] inferred two estimates of divergence time between *C. elegans* and *C. briggsae*. One was 89±13 Myr, and the other was 114±22 Myr. To provide a rough prior on the substitution rate for our data set, we used the BASEML program of the PAML package[12] to estimate the branch lengths and

Table 2.1: Divergence times for the primate data set analyzed under the autocorrelated lognormal relaxed clock. The posterior means and 95% CIs of divergence times (Myr), mean rates and rate of the human branch (expected number of substitutions per site per million years) are reported. Divergence times are defined in Figure 2.1.

	Prior	Prior & Mutation data (Generation time: 20 years)	Sequence data	Sequence data & Mutation data (Generation time: 20 years)	Sequence data Mutation data (Generation time: 30 years)
t_1(root)	41.26	41.21	40.88	45.09	45.97
	(30.36,53.12)	(30.02,52.99)	(29.91,52.38)	(34.06,57.04)	(34.23,57.59)
t_2	25.19	25.18	23.38	37.03	39.3
	(5.7,44.43)	(5.22,44.03)	(11.58,36.17)	(24.54,50.18)	(25.48,52.48)
t_3	13.36	13.44	13.54	24.28	27.71
	(0.13,30.33)	(0.35,30.82)	(6.46,22.19)	(15.89,33.42)	(17.91,38.27)
t_4	5.65	5.61	4.15	8.62	11.53
	(0,17.86)	(0,17.9)	(1.65,7.27)	(5.05,12.81)	(6.58,17.47)
t_5	9.52	9.46	4.42	7.34	8.3
	(0,26.7)	(0,26.97)	(1.83,7.31)	(3.2,11.84)	(3.4,14.73)
mean rate	0.0014	0.00086	0.0016	0.0012	0.00112
	(0,0.0046)	(0.00012,0.0021)	(0.0011,0.0021)	(0.0009,0.0015)	(0.0008,0.00144)
human rate	0.0014	0.00054	0.0016	0.00056	0.00037
	(0,0.005)	(0.00044,0.00064)	(0.0006,0.0025)	(0.00046,0.00066)	(0.00031,0.00044)

the substitution rates for the fourfold degenerate sites of the 21 genes under the strict clock that was calibrated by a time of 101.5 Myr (median of 89 Myr and 114 Myr) for the most recent common ancestor between *C. elegans* and *C. briggsae*. The prior on the substitution rate (for the strict clock) or the root rate (for the autocorrelated lognormal clock) was therefore a gamma distribution with mean 0.008 and standard deviation 0.004 substitutions per site per million years. The prior for the variance parameter ν in the autocorrelated clock was gamma with mean 0.01 and standard deviation 0.01. The prior for the root time was specified through the tree prior, which is a birth-death process. The priors for κ, α and the parameters in the birth-death process were the same as the previous primates case. All loci in the analysis were assumed to share the same clock, substitution and tree models.

The posterior means and 95% CIs for the divergence times under the strict clock and the autocorrelated lognormal relaxed clock for generation times of 14, 30 and 60 days are shown in Table 2.2 and Table 2.3. The inferred divergence date of *C. elegans* and *C. briggsae* with the mutation data was substantially more recent than the results from Blair et al.[145], but it was earlier than the dates that Cutter reported[112].

Cutter estimated the branch-specific synonymous-site divergence (d_s) as a proxy for the neutral divergence and converted d_s to the divergence time by a calibration of the mutation rate under the molecular clock assumption[112]. But in our study, instead of doing a two-step procedure, we modeled the mutation data directly and allowed the likelihood of the mutation data to contribute to divergence time estimation. The discrepancy between ours and Cutter's

results[112] might partially be attributed to the fact that we only analyzed 21 of the loci that they collected and we only analyzed the fourfold degenerate sites from these 21 loci. Moreover, in addition to the strict clock, our method allowed the clock assumption to be relaxed so that data were analyzed under the autocorrelated relaxed clock as well. It is worth mentioning that the mutation data have a dramatic effect on the divergence times when no calibrations of times are available. Unlike the primate data set, the prior on the root time in this study was indirectly specified through the birth-death tree prior instead of using a gamma prior to indicate the expected date. Hence, the marginal prior on the root time was close to zero (results not shown). In this case, the divergence times were strongly influenced by the calibration from the mutation data and the change of the generation times made an enormous difference. Therefore, we reached a conclusion similar to Obbard et al.[3] that mutation data might be more useful for dating divergence times when more data (e.g. generation time in the wild) and better models are available.

Table 2.2: Divergence times for the nematode data set analyzed under the strict clock. The posterior means and 95% CIs of divergence times (Myr) are reported. Divergence times are defined in Figure 2.2(a). Three different mean generation times of *C. elegans* are assumed, 60, 30, and 14 days.

	Sequence data	Sequence & Mutation data (Generation time: 60 days)	Sequence & Mutation data (Generation time: 30 days)	Sequence & Mutation data (Generation time: 14 days)
t_1(root)	142.4 (40.03,294.13)	70.11 (58.19,83.08)	35.77 (29.63,42.33)	17.47 (14.43,20.66)
t_2	109.33 (32.49,227.1)	53.87 (45.68,62.58)	27.47 (23.24,31.84)	13.42 (11.39,15.57)
t_3	95.12 (28.05,197.02)	46.86 (40.05,54.03)	23.91 (20.47,27.58)	11.67 (9.97,13.44)
t_4	81.84 (23.95,169.39)	40.33 (34.67,46.26)	20.57 (17.76,23.62)	10.05 (8.67,11.55)
t_5	75.9 (22.05,157.58)	37.41 (31.9,43.17)	19.08 (16.25,22.03)	9.32 (7.91,10.73)

Table 2.3: Divergence times for the nematode data set analyzed under the autocorrelated log-normal relaxed clock. Posterior means and 95% CIs of divergence times (Myr) are reported. Divergence times are defined in Figure 2.2(a). Three different mean generation times of *C. elegans* are assumed, 60, 30, and 14 days.

	Sequence data	Sequence & Mutation data (Generation time: 60 days)	Sequence & Mutation data (Generation time: 30 days)	Sequence & Mutation data (Generation time: 14 days)
t_1(root)	115.04 (34.88,234.02)	67.43 (43.97,91.86)	42.93 (25.41,64.18)	28.55 (17.31,40.64)
t_2	88.22 (23.54,181.22)	49.67 (37.99,61.31)	24.85 (18.53,31.52)	11.34 (8.43,14.4)
t_3	71.23 (17.67,146.5)	40.91 (26.72,53.93)	20.21 (12.31,27.91)	8.94 (5.45,12.72)
t_4	60.47 (14.97,125.36)	34.78 (22.11,46.42)	17.04 (9.94,23.87)	7.44 (4.36,10.77)
t_5	55.28 (13.84,115.00)	31.83 (19.76,42.83)	15.51 (8.93,22.07)	6.71 (3.8,9.85)

2.3.3 Divergence Times of Fruit Flies

The results from Obbard et al.[3] suggested that there was between-branch rate variation in 12 *Drosophila* sequences, so we only analyzed the dataset under the autocorrelated lognormal relaxed clock. In order to specify the prior on the root rate, a calibration of 5.4 Myr for the most recent common ancestor of *D. melanogaster* and *D. simulans* (Tamura et al.[2]) was selected. We estimated a substitution rate of 0.016 changes per site per million years under the strict clock by BASEML program in PAML package[12]. It led to a gamma prior on the root rate with mean 0.016 and standard deviation 0.008 per site per million years. The inferred root time under the strict clock was 50 Myr. Therefore, as suggested by Kishino et al.[10], the prior for the variance parameter ν in the autocorrelated clock was gamma with mean $1/50 = 0.02$ and standard deviation 0.02. The prior for the times were the same as the previous nematodes dataset, a birth-death prior with the default priors on the birth and the relative death rates in BEAST[14].

The rooted tree topologies were different with or without the mutation data (Figure 2.4), but they shared the same unrooted tree topology in Figure 2.3. Since the rate on each branch was sampled, BEAST[14] can sample the root position and produce a rooted phylogeny. In our case, the mutation data from *D. melanogaster* provided information for the rates along the branch, and therefore it had an effect on the root placement. Similarly, since different clock models inferred the rates on the branch differently, alternative root positions were selected. In fact, the derived root positions were different from Obbard et al.[3]. And neither Obbard et al.[3] nor our results resemble the widely accepted rooted tree topology from Clark et al.[7] in Figure 2.2(b). We inferred the root position on Branch A if mutation data were used and on Branch B if not, while Obbard et al. placed the root on Branch C in Figure 2.2(b).

The posterior sample of trees under the autocorrelated clock was summarized by TreeLog-Analyser in BEAST[14]. There were 10 unique clades (including the clade that contains all 12 species) supported by 80% or above of the posterior tree samples for both the sequence data only and with the mutation data case (see Figure 2.4 for the posterior support values). All other clades were supported by less than 50% of the posterior tree samples. As suggested by Obbard et al.[3], there was little information from only 12 species to allow estimation of relaxed-clock parameters. Therefore, the uncertainty associated with the rates and the times was too large so that BEAST[14] could not root the tree properly.

The posterior means and 95% CIs of the divergence times are shown in Table 2.4. Our divergence time estimates of Nodes 1 - 4 with the *D. melanogaster* mutation data resembled those Obbard et al.[3]. However, for Nodes 5 - 7, the inferred divergences and the widths of the 95% credible interval differed. In the case without proper calibrations of times, the closer the nodes were to the *D. melanogaster*, the narrower the widths of the 95% credible

interval would be under the autocorrelated lognormal relaxed clock model. This is because only the mutation data and the prior could provide information to separate rates and times, and the effect of the *D. melanogaster* mutation data attenuated for farther apart branches. In contrast, the uncorrelated lognormal relaxed clock, employed by Obbard et al.[3], utilized the *D. melanogaster* mutation rate to specify the mean of a lognormal distribution to sample branch rates. In other words, the *D. melanogaster* mutation rate can be viewed as a rough calibration for all the branch rates. Therefore, the widths of the 95% credible interval under the uncorrelated lognormal relaxed clock were not as wide as under the autocorrelated lognormal relaxed clock for divergence times farther away from *D. melanogaster*. However, whether all 12 *Drosophila* species shared similar mutation rates is unjustified and this assumption did have impact on divergence time estimation.

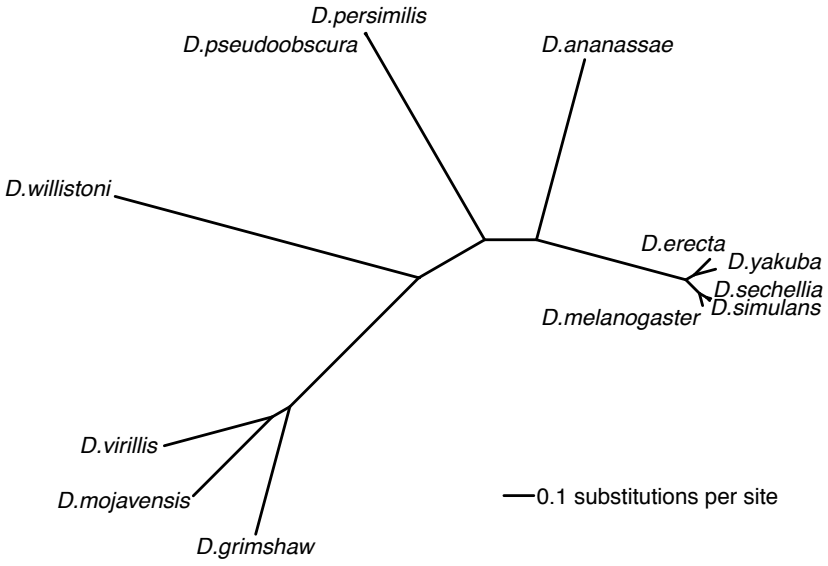


Figure 2.3: Unrooted phylogram of 12 species of *Drosophila*.

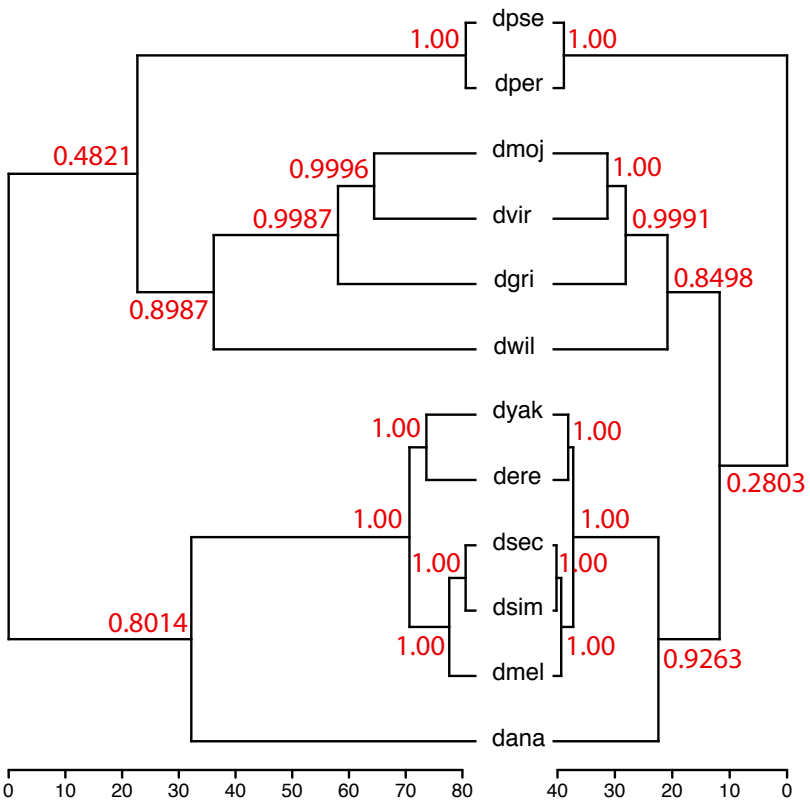


Figure 2.4: The phylogeny of 12 species of *Drosophila* inferred under an autocorrelated log-normal relaxed clock. The node times are the posterior means. Each node is labeled with its posterior support value. The left-hand tree was inferred without using laboratory estimated neutral mutation rate. The right-hand tree was inferred with mutation information from *D. melanogaster*. Note that the two phylogenies root the tree differently but they share the same unrooted tree. Species abbreviations: dpse, *D. pseudoobscura*; dper, *D. persimilis*; dmoj, *D. mojavensis*; dvir, *D. virilis*; dgri, *D. grimshawi*; dwil, *D. willistoni*; dyak, *D. yakuba*; dere, *D. erecta*; dsec, *D. sechellia*; dsim, *D. simulans*; dmel, *D. melanogaster*; dana, *D. annanassae*.

2.3.4 Limitations and Conclusions

This study provides a method to incorporate contemporary mutation data into the interspecific divergence time estimation problem. In this study, instead of treating information from the mutation data as prior knowledge, the contemporary mutation data act as a second source of data that has its own likelihood and directly influences the posterior distribution of the divergence times. The uncertainty in the experimentally derived mutation rate estimates is incorporated in the mutation model. Rather than using the mutation data to calibrate the rate of every branch on the tree, our treatment only applies the mutation data to the terminal

Table 2.4: Divergence times (Myr) under the autocorrelated lognormal clock for the 12 species of *Drosophila* data set. The number in parentheses after the taxa compared is the node number in Figure 2.2(b). Results from Russo et al.[1], Tamura et al.[2] and Obbard et al.[3] are represented for comparison. Ten generations per year were assumed to convert the mutation rate per site per generation to the mutation rate per site per million years for Obbard et al.[3] and our studies.

Taxa compared	Previously published estimates			Sequence data	Sequence & Mutation data
	Russo et al. [1]	Tamura et al. [2]	Obbard et al. [3]		
simulans complex (1)	-	-	0.5 (0.3,0.7)	1.19 (0.11,2.92)	0.48 (0.36,0.61)
melanogaster-simulans (2)	2.3 (1.02,3.57)	5.4 (3.24-7.56)	1.4 (0.9,1.9)	3.25 (0.41,8.09)	1.3 (1.06,1.55)
yakuba-erecta (3)	-	10.4 (5.89,14.91)	2.4 (1.6,3.3)	6.31 (0.71,15.92)	2.5 (1.78,3.31)
melanogaster subgroup (4)	6.1 (3.9,8.3)	12.8 (7.51,18.09)	3.3 (2.4,4.4)	8.03 (0.98,19.91)	3.39 (2.65,4.21)
melanogaster-annanassae (5)	-	44.2 (26.76,61.64)	15 (11,21)	37.01 (4.72,90.43)	18.59 (13.2,24.83)
virillis/repleta-hawaiian (6)	32.2 (26.24,38.16)	42.9 (25.85,59.95)	13 (9.6,17)	18.49 (1.99,44.96)	12.51 (2.66,26.24)
melanogaster-willistoni (7)	36.3 (27.95,44.64)	62.2 (27.31,87.09)	32 (25,40)	59.68 (13.97,133.28)	35.88 (16.41,57.09)

branches. This not only allows the mutation information from one species to influence other branch rates, but it also takes into account the attenuation of rate information with time. From the empirical data analyses conducted above, the mutation data have a great impact on divergence time estimation. Mutation data can even influence the rooting of the inferred tree topology as demonstrated in the *Drosophila* data analysis.

This method should be more useful for dating a group of species with poor fossil records, but this method should also be used with caution. Without a treatment of natural selection, the mutation data can only be applied to neutrally evolving regions. Fortunately, incorporation of mutation-selection balance into models of molecular evolution is becoming increasingly common. These models (e.g. Halpern and Bruno[115]) considering natural selection and could be used to derive substitution rates from mutation data. Our application of mutation data relies on the assumption of a constant rate or at least autocorrelated rates. It is unlikely that mutation data will prove very helpful for divergence time estimation if substitution rates between branches are uncorrelated. The potential value of mutation data for divergence time estimation is tightly connected to how substitution rates change over time. In general, mutation information from the tips of a phylogeny will be more helpful for estimating evolutionary rates on branches near these tips. Mutation data will be less helpful for illuminating rates of branches that are far from tips, especially if rate autocorrelation is weak.

This study is meant to be exploratory; it is not intended to be applied to all regions in the genomes, but it can be generalized. Some of the generalizations are described in Chapter 5.

Chapter 3

The impact of the mutation data on Bayesian divergence time estimation

3.1 Introduction

Because the true divergence times are always unknown in the real world, it is hard to quantify the influence of the mutation data on Bayesian divergence time estimation. In this chapter, the impact of the mutation data on divergence time estimation is examined analytically and with simulations. In addition, we study the limit of divergence time estimation when the number of sites in the mutation data approaches infinity. We reach a similar conclusion to that of Rannala and Yang's "infinite-site theory"[47]. This uncertainty in times does not go to zero even if mutation data are available for all extant species so that all terminal branch rates are known. Moreover, infinite sequence data as well as infinite mutation data can not reduce the errors in posterior time estimates to zero unless with a very simple case.

3.2 Inference with Infinite Data

3.2.1 Infinite Mutation Data with the Strict Molecular Clock

Before considering the autocorrelated relaxed clock, we first consider the situation for a constant substitution rate. We will assume the mean generation time g is known and will simplify notation by omitting parameters Ω and Θ. The posterior distribution of rate and times is

$$f(\mathbf{t}, r|\mathbf{X}, \mathbf{M}) = \frac{f(\mathbf{X}|\mathbf{t}, r)f(\mathbf{M}|r)f(\mathbf{t}, r)}{f(\mathbf{X}, \mathbf{M})} = \frac{f(\mathbf{X}|\mathbf{t}, r)f(\mathbf{M}|r)f_T(\mathbf{t})f_R(r)}{f(\mathbf{X}, \mathbf{M})}, \tag{3.1}$$

where $f_T(\mathbf{t})$ and $f_R(r)$ are the priors for times and rate, respectively. Note that for the strict molecular clock case, there is only one rate for the tree and therefore r is a scalar. To demonstrate, assume an improper prior for r, with $f_R(r) = 1$ for $0 < r < \infty$.

For infinite mutation data, the true rate $r = r^*$ is assumed known without error. Then the posterior for times $\mathbf{t}$ given the alignment $\mathbf{X}$ and the true rate r is

$$\begin{aligned} f(\mathbf{t}|\mathbf{X}, r = r^*) &= \frac{f(\mathbf{t}, r = r^*, \mathbf{X})}{\int f(\mathbf{t}, r = r^*, \mathbf{X}) d\mathbf{t}} = \frac{f(\mathbf{X}|\mathbf{t}, r = r^*) f_R(r^*) f_T(\mathbf{t})}{\int f(\mathbf{X}|\mathbf{t}, r = r^*) f_R(r^*) f_T(\mathbf{t}) d\mathbf{t}} \\ &= \frac{f(\mathbf{X}|\mathbf{t}, r = r^*) f_T(\mathbf{t})}{\int f(\mathbf{X}|\mathbf{t}, r = r^*) f_T(\mathbf{t}) d\mathbf{t}}. \end{aligned} \tag{3.2}$$

This formula implies that even when the amount of mutation data increases to infinity, the posterior for times does not converge to a degenerate distribution with values equal to the true divergence times. Instead, uncertainty in times remains and the amount of uncertainty depends on the aligned sequences and the true value of the rate.

By comparing Equation (3.2) with the marginal posterior distribution of time $\mathbf{t}$, which is

$$f(\mathbf{t}|\mathbf{X}) = \int_0^\infty f(\mathbf{t}, r|\mathbf{X}) dr = \frac{\int_0^\infty f(\mathbf{X}|\mathbf{t}, r) f_R(r) f_T(\mathbf{t}) dr}{\int\int f(\mathbf{X}|\mathbf{t}, r) f_R(r) f_T(\mathbf{t}) dr d\mathbf{t}} = \frac{\left(\int_0^\infty f(\mathbf{X}|\mathbf{t}, r) dr\right) f_T(\mathbf{t})}{\int\int f(\mathbf{X}|\mathbf{t}, r) f_T(\mathbf{t}) dr d\mathbf{t}}, \tag{3.3}$$

we see that whether infinite mutation data can reduce the uncertainty in times depends on the difference between the likelihood of the aligned sequences when the rate is known $f(\mathbf{X}|\mathbf{t}, r = r^*)$ and when considering all possible rates $\int f(\mathbf{X}|\mathbf{t}, r) dr$. For simplicity, consider only the root time t and omit the notation of aligned sequences $\mathbf{X}$ for now. In the discussion below, the distribution of time t is conditioned upon the aligned sequence data. That is, the variance of t also depends on the branch length. The conditional variance of t given the rate is a random variable, and the value of $Var(t|r)$ depends on the value of rate r.

According to the Law of total variance,

$$Var(t) = Var\left(E(t|r)\right) + E\left(Var(t|r)\right). \tag{3.4}$$

Since variance is always non-negative, $Var\left(E(t|r)\right)$ is non-negative. Therefore, Equation (3.4) implies that $Var(t) \geq E\left(Var(t|r)\right)$, where $E\left(Var(t|r)\right) = E\left(c(r)\right)$ can be viewed as a weighted average of the conditional variance $Var(t|r)$ and where $c(r)$ is a function of the rate r. The expected value $E\left(Var(t|r)\right)$ can be either greater or less than $Var(t|r)$ depending on the value of r, and whether the marginal variance $Var(t)$ is greater than conditional variance $Var(t|r)$ is determined by the value of r as well. In other words, the conditional variance $Var(t|r)$ is not guaranteed to be smaller or equal to the marginal variance $Var(t)$.

We illustrate this with a simple example. Consider a case of 2 species evolving under the strict clock. Under the Jukes and Cantor model, data are summarized with x differences in n sites in a pairwise alignment[49]. The likelihood of observing the sequence alignment given rate r and time t is

$$L(r,t|n,x) = \binom{n}{x}\left(\frac{3}{4}-\frac{3}{4}e^{-8rt/3}\right)^{x}\left(\frac{1}{4}+\frac{3}{4}e^{-8rt/3}\right)^{n-x}. \tag{3.5}$$

Suppose there are 55 differences in 100 sites in the pairwise alignment. The likelihood function is plotted in Figure 3.1. It shows that if the true rate is high (e.g. $r = 2$), the uncertainty in time is relatively small. In contrast, when the true rate is low (e.g. $r = 0.5$), then the uncertainty in time is rather large. While the posterior distribution of time will also depend on its prior distribution, our point is that having a great amount of mutation data is not guaranteed to yield less uncertainty about time. The uncertainty in time depends on prior information, the rate information in the mutation data, and on the branch length information from the interspecific sequence data.

Finally, when the amount of both molecular sequence data and mutation data increases to infinity, we can estimate the branch length $\mathbf{b}$ and rate r without error. And since the branch length is the product of rate and times, $\mathbf{b} = r\mathbf{t}$, we can also estimate the divergence times without error.

3.2.2 Infinite Mutation Data with an Autocorrelated Lognormal Relaxed Clock

Several models for autocorrelated substitution rates assume the rate follows a probability distribution with mean determined by the ancestral rate and with variance depending on the time separating the ancestral and descendant rates. In addition to a variance parameter ν for specifying the degree of autocorrelation, the rate at the root r_0 has to be treated differently and assigned its own prior $f(r_0)$. Then given the tree topology τ, the prior distribution for rates $f(\mathbf{r}|\mathbf{t})$ is simply the product of the probabilities of individual rates r_i given ancestral rates $r_{A(i)}$ and the time separating the ancestral and derived rate,

$$f_R(r_0,\mathbf{r}|\mathbf{t}) = f(r_0)\prod_{i=1} f(r_i|r_{A(i)},\mathbf{t}). \tag{3.6}$$

Although several previously proposed treatments of autocorrelated substitution rates have forms described by Equation (3.6), their details differ. Here, we briefly summarize three quite similar models for autocorrelated rates.

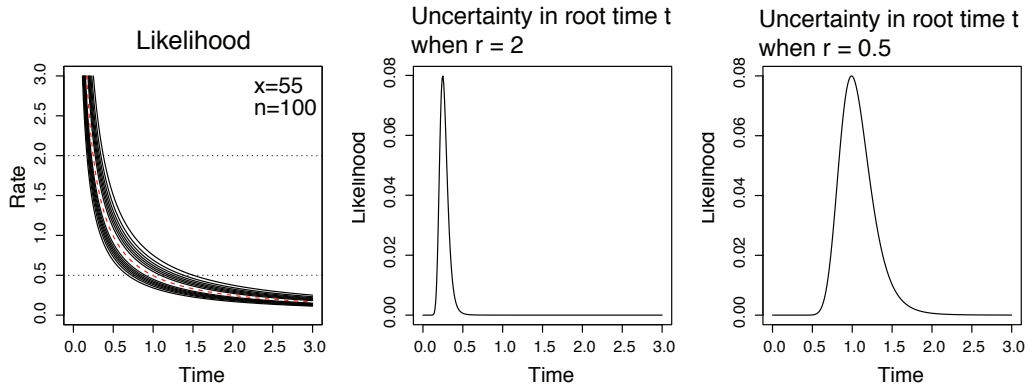


Figure 3.1: Likelihood of the pairwise alignment with $x = 55$ differences in $n = 100$ sites. The maximum likelihood estimate of the branch length is the red dashed line in the left figure. The central figure represents the likelihood function conditioned on a high rate ($r = 2$). The right figure represents the likelihood function conditioned on a low rate ($r = 0.5$). These figures show that the uncertainty of the root time can be very different when conditioned on different rates.

Autocorrelated Rates for Nodes Kishino et al.[10] proposed a geometric Brownian motion process to model the rates for nodes along lineages. In their model, the logarithm of the rate at one node has a normal distribution given the logarithm of the rate at its ancestral node. Let us denote the rate at the node that begins a branch as r_A and the rate at the node ending the branch as r with time duration t of the branch. Then by forcing the expected value of r to be r_A, they modeled

$$\log(r)|\log(r_A) \sim N\left(\log(r_A) - \frac{t\nu}{2}, t\nu\right). \tag{3.7}$$

The variance parameter, ν, is a non-negative real number, and it governs the degree of autocorrelation. A strict clock occurs when ν equals zero. Small values of ν represent stronger autocorrelation than large values. Kishino et al.[10] treat the rate on a branch as the average of the rates at the nodes at its beginning and ending. The branch length b for branch starting with rate r_A and ending with rate r is therefore $b = (r_A + r)t/2$.

Autocorrelated Rates for Midpoints of Branches Instead of modeling the rate at nodes, Rannala and Yang[47] modeled the rate at the midpoint of a branch and used it as a proxy for the rate for the branch. Let r'_A be the rate at the midpoint of a parent branch with total time duration $2t'_A$ and let r' be the rate at the midpoint of the daughter branch with time duration of the branch $2t'$. Rannala and Yang[47] had

$$\log(r')|\log(r'_A) \sim N\left(\log(r'_A) - \frac{(t'_A + t')\nu}{2}, (t'_A + t')\nu\right). \tag{3.8}$$

For the rates of the descendant branches of the root, they plugged in $r'_A = r_0$ and $t'_A = 0$ in Equation (3.8). In this parametrization, the branch length b of the branch with time duration t and rate at the midpoint r' is simply $b = r' \times t$.

Autocorrelated Rates for Branches The authors of BEAST[14] described an earlier version of their autocorrelated lognormal relaxed clock (ACLN) model in Ho et al.[146] and in Drummond et al.[46]. In this earlier version, the rate in each branch is drawn from a lognormal distribution whose mean is equal the rate in its parent branch. They adopted the formula in Rannala and Yang[47] and implemented a later version of the autocorrelated lognormal relaxed clock (ACLikelihood class) in BEAST. However, this later variant directly models the average rate over each branch but not the rate of midpoint of each branch. We will refer to the later version of the ACLN model in this paper.

Let $\bar{r}$ and $\bar{r}_A$ be the average rates of the descendant and ancestral branches, respectively. Let t be the time duration of the descendant branch. Then the ACLN model has

$$\log(\bar{r})|\log(\bar{r}_A) \sim N\left(\log(\bar{r}_A) - \frac{t\nu}{2}, t\nu\right). \tag{3.9}$$

Note that the variance of the lognormal distribution is proportional to the time duration of the descendant branch, but is independent of the time duration of the ancestral branch. For the daughter branches of the root, BEAST sets the ancestral branch rate as the root rate $\bar{r}_A = r_0$ in Equation (3.9). Finally, the branch length b of a branch with average rate along the branch $\bar{r}$ and time duration t is $b = \bar{r} \times t$.

Bayesian Framework

For simplicity, consider the parametrization of the ACLN model in BEAST[14] and assume that every species has infinite mutation data with known generation times. In this parameterization, the rate of evolution $\mathbf{r}$ is a vector of branch rates. The likelihood of the sequence data depends on the branch lengths of the unrooted tree in Figure 3.2(a). Because the root is not specified, the root rate r_0 is neglected in the discussion below. We assume that infinite mutation data are

available for every extant species so that the rate for each terminal branch $\mathbf{r}_T = \mathbf{r}_T^*$ is known without error.

Note that originally there are $3s-3$ parameters in the ACLN model: $s-1$ divergence times, $2s-2$ branch rates. Knowing the rates for all terminal branches reduces the dimension by s, so that there are $2s-3$ parameters remaining. If, however, the number of sites in the alignment goes to infinity as well so that the branch lengths in the unrooted tree can be treated as observed data, then the dimension in the parameter space can be further reduced. The number of parameters left to be estimated depends on the tree topology. For example, consider the two different tree topologies with four species that are shown in Figure 3.2. Suppose that there is infinite mutation data from all species on the tree so that all terminal branch rates $\{r_1, r_2, r_3, r_4\}$ can be estimated without error. Then there are still $3s-3-s=5$ parameters left to be estimated for the two tree topologies: $\{r_5, r_6, t_5, t_6, t_7\}$. On the other hand, if infinite molecular sequence data are also available so that the branch lengths in Figure 3.2(a) can be treated as known, then the number of remaining parameters in the model are not the same for the balanced tree (Figure 3.2(b)) and the unbalanced tree (Figure 3.2(c)). For the tree in Figure 3.2(b), the divergence times are expressed by

$$t_5 = b_1/r_1 = b_2/r_2 \tag{3.10}$$

$$t_6 = b_3/r_3 = b_4/r_4 \tag{3.11}$$

$$(t_7 - t_6) \times r_6 + (t_7 - t_5) \times r_5 = b_5 \Rightarrow t_7 = \frac{b_5 + t_5 r_5 + t_6 r_6}{r_5 + r_6}. \tag{3.12}$$

Therefore, t_5 and t_6 can be analytically determined by the branch lengths and the terminal branch rates, and t_7 depends on the values of r_5 and r_6. There are only 2 free parameters left in the model: $\{r_5, r_6\}$. Note that either r_1 or r_2 alone can determine t_5 so that the infinite mutation data on Tips 1 and 2 are actually redundant. Similarly, infinite mutation data on Tips 3 and 4 are redundant. These infinite amount of mutation data on tips do not directly relate to the root time t_7.

In contrast, for the tree in Figure 3.2(c), the divergence times can be expressed by

$$t_5 = b_1/r_1 = b_2/r_2 \tag{3.13}$$

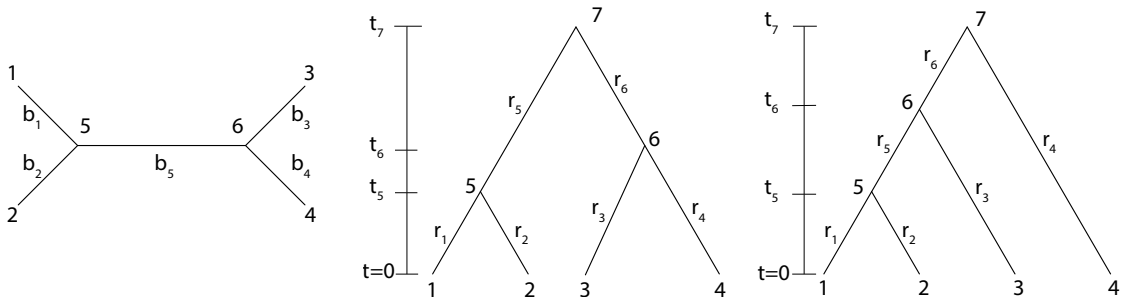
$$t_6 = b_3/r_3 = t_5 + b_5/r_5 \Rightarrow r_5 = \frac{b_5}{t_6 - t_5} = \frac{b_5}{b_3/r_3 - b_1/r_1} \tag{3.14}$$

$$(t_7 - t_6) \times r_6 + t_7 \times r_4 = b_4 \Rightarrow t_7 = \frac{b_4 + t_6 r_6}{r_4 + r_6}. \tag{3.15}$$

In this case, not only t_5 and t_6 but also r_5 can be determined by the branch lengths and the terminal branch rates. Moreover, once r_6 is fixed, t_7 can be determined as well. In other words

and in contrast to the case of the balanced tree, only 1 free parameter remains. Even though the infinite mutation data on Tips 1 and 2 are redundant, the infinite mutation data on Tip 3 and Tip 4 can help solve t_6 and t_7, respectively.

The shape of the underlying phylogeny can influence the dimensionality of parameter space when infinite sequence and mutation data are present, and therefore can have an impact on the uncertainty of divergence time estimation. Note that the above discussion is based on the assumption that the mutation rate at the tip is equal to the average rate on its branch. Without this assumption, the terminal branch rates remain uncertain and both tree topologies have $2s-3$ parameters in the parameter space. In this case, mutation data on the tips influence the divergence time estimation through reducing the uncertainty in terminal branch rates.



(a) Unrooted tree topology (b) Balanced tree (c) Unbalanced tree

Figure 3.2: Two rooted trees with the same unrooted tree topology. The b's are the branch lengths in the unrooted tree. The r's are the branch rates in the rooted tree.

Omitting parameters Ω and Θ for the sake of simplicity, the posterior distribution of branch rates and times given the mutation data and molecular sequence data is

$$f(\mathbf{t},\mathbf{r}|\mathbf{X},\mathbf{M}) = \frac{f(\mathbf{X}|\mathbf{t},\mathbf{r})f(\mathbf{M}|\mathbf{r})f(\mathbf{t},\mathbf{r})}{f(\mathbf{X},\mathbf{M})} = \frac{f(\mathbf{X}|\mathbf{t},\mathbf{r})f(\mathbf{M}|\mathbf{r}_T)f_R(\mathbf{r}|\mathbf{t})f_T(\mathbf{t})}{f(\mathbf{X},\mathbf{M})}, \tag{3.16}$$

where $f_R(\mathbf{r}|\mathbf{t})$ is the prior for branch rates given the times with the form as in Equation (3.6).

With infinite mutation data for all taxa, the posterior distribution for times, internal branch

rates and root rate becomes

$$f(\mathbf{t}, \mathbf{r}_{-T}, r_0|\mathbf{X}, \mathbf{r}_T = \mathbf{r}_T^*) = \frac{f(\mathbf{t}, r_0, \mathbf{r}_{-T}, \mathbf{r}_T = \mathbf{r}_T^*, \mathbf{X})}{f(\mathbf{X}, \mathbf{r}_T)} \tag{3.17}$$

$$= \frac{f(\mathbf{X}|\mathbf{r}_{-T}, \mathbf{r}_T = \mathbf{r}_T^*, \mathbf{t}) f_R(r_0, \mathbf{r}_{-T}, \mathbf{r}_T = \mathbf{r}_T^*|\mathbf{t}) f_T(\mathbf{t})}{\int\int f(\mathbf{t}, r_0, \mathbf{r}_{-T}, \mathbf{r}_T = \mathbf{r}_T^*, \mathbf{X}) d\mathbf{t} d\mathbf{r}_{-T}} \tag{3.18}$$

Equation (3.18) implies that, similar to the strict clock case, uncertainty in times and rates at the internal nodes does not vanish even if infinite mutation data are available for all species analyzed. The first term in the numerator in Equation (3.18), $f(\mathbf{X}|\mathbf{r}_{-T}, \mathbf{r}_T = \mathbf{r}_T^*, \mathbf{t})$, suggests that there is an identifiability problem. In fact, even if all branch rates can be estimated without error as in the strict clock case, uncertainty in times remains because branch length is the product of rate and time. Moreover, if we have a large amount of molecular sequence data in addition to infinite mutation data so that we can estimate all terminal branch lengths without error as well, there are still some internal branches with uncertainty in branch rates where we are unable to disentangle rates and times. The situation is similar to the strict clock case, in which the variance in times may or may not be reduced by conditioning on the terminal branch rates.

On the other hand, the joint prior for rates given times, $f_R(r_0, \mathbf{r}_{-T}, \mathbf{r}_T = \mathbf{r}_T^*|\mathbf{t})$ as in the second term in the numerator in Equation (3.18), contributes to uncertainty in internal branch rates as well. It can be decomposed into a form different than Equation (3.6) as follows.

$$f_R(r_0, \mathbf{r}_{-T}, \mathbf{r}_T = \mathbf{r}_T^*|\mathbf{t}) = f(\mathbf{r}_{-T}|\mathbf{r}_T = \mathbf{r}_T^*, r_0, \mathbf{t}) f(\mathbf{r}_T = \mathbf{r}_T^*|r_0, \mathbf{t}) f(r_0). \tag{3.19}$$

The first term in Equation (3.19) is the prior of internal branch rates conditional on the end-points (terminal branch rates and the rate at the root) and times. And the second term is the probability of terminal branch rates given the root rate and times, which is a constant and can be easily calculated by plugging the true terminal branch rates, the root rate and the times into the ACLN model. The last term is the prior of root rate, and we write $f(r_0)$ instead of $f(r_0|\mathbf{t})$ by assuming that the rate at the root is independent of divergence times $\mathbf{t}$. For the posterior distribution of times only, we need to integrate out all possible values of root rate r_0 and internal branch rates $\mathbf{r}_{-T}$ in Equation (3.18).

If the rates along the lineages are described by a geometric Brownian process (so the logarithm of rates are described by Brownian motion), then the process becomes a geometric Brownian bridge when the rates at the ancestral and the descendant branches are "tied down" to specific values. In this case, the variance of internal branch rates $\mathbf{r}_{-T}$ can be reduced by conditioning on the terminal branch rates $\mathbf{r}_T$. In other words, information about internal branch rates is provided by knowing the terminal branch rates, and hence uncertainty in internal branch

rates can be reduced.

Brownian bridge example Here we illustrate how infinite mutation data on the tips will reduce the variance associated with the internal rates. Without loss of generality, let us define $\{B(t); t \geq t_0\}$ as a Brownian motion process starting at $B(t_0) = x$ with variance parameter ν ($\nu > 0$). Then the expected value and the variance of $B(t)$ are $E[B(t)] = x$ and $Var[B(t)] = \nu(t-t_0)$, respectively. By conditioning on the event that the values at the beginning and the end are restricted to specific values $\{B(t_0) = x, B(t_1) = y, t_1 > t_0\}$, we can construct a Brownian bridge $\{B^0(t); t \geq t_0\}$. For $t_0 < t < t_1$, we can derive the expected value and the variance of the Brownian bridge $B^0(t)$, which are $E[B^0(t)] = x + \frac{t-t_0}{t_1-t_0}(y-x)$ and $Var[B^0(t)] = \frac{\nu(t_1-t)(t-t_0)}{t_1-t_0}$ (see Guindon[147] for derivations). By taking the differences of the variances with or without conditioning on the end point $B(t_1) = y$, we get

$$Var[B(t)] - Var[B^0(t)] = \nu(t-t_0) - \frac{\nu(t_1-t)(t-t_0)}{t_1-t_0} = \frac{\nu(t-t_0)^2}{t_1-t_0} > 0. \tag{3.20}$$

In other words, the variance of Brownian motion is reduced if we conditioned on the end points. It implies that if the rate evolves like a Brownian motion process and there is infinite amount of mutation data for a species so the rate of that species is known, then the variance of the rate along the branch is reduced.

3.3 Analytical Results for Two Species Case

3.3.1 Strict Clock

We will consider a case of 2 species evolving under the strict clock with the same settings as Dos Reis and Yang[148]. We assume the interspecific sequence data are generated under Jukes and Cantor (1969) model[49]. Therefore, the data are summarized with x differences in n sites in a pairwise alignment[49]. We set the divergence time to be $t = 1$ and the rate to be $r = 0.5$. Then the true branch length, or the expected number of substitutions per site between the two sequences is $b = r \times 2t = 1$. If one time unit is one million years (My), then the chronological rate of evolution is 10^{-6} substitutions per site per year.

Consider a gamma prior such that the rate per million years has mean 0.5 and variance 0.125 (i.e. $r \sim G(2,4)$), and a gamma prior on the time with mean 1 and variance 0.5 (i.e. $t \sim G(2,2)$) (Figure 3.3). This means the joint prior of rate and time is

$$f_{RT}(r,t) = f_R(r)f_T(t) = 16re^{-4r} \times 4te^{-2t}. \tag{3.21}$$

The likelihood of observing the sequence alignment given rate r and time t is

$$L(r,t|n,x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \left(\frac{3}{4} - \frac{3}{4}e^{-8rt/3}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-8rt/3}\right)^{n-x}. \quad (3.22)$$

With this simple case, we can illustrate the influence of alignment length and amount of mutation data on inference of rate and time. Consider two alignment lengths. Both have the observed proportion of differences being $\frac{x}{n} = 0.55$ but one has $n = 100$ and the other has $n = 1000$. We also consider two mutation data sets. Each contains y observed mutations in a total m bases assayed in one generation. Both sets have the proportion of mutations among bases assayed being $\frac{m}{y} = 10^{-5}$, but the smaller mutation data set has $m = 10^6, y = 10$ and the bigger has $m = 10^7, y = 100$ with a generation time being 10 years. For each of the mutation data sets, m is considered to be a realization from a Poisson random variable with mean $10^{-6} \times 10 = 10^{-5}$ substitutions per site per generation. The likelihood function is shown in Figure 3.4.

Combining the rate information from the mutation data with the prior knowledge of rate and time, the uncertainty in rate can be substantially reduced. The uncertainty is further reduced if more mutation data are available (See Figure 3.3). However, the uncertainty in times remains the same regardless of the increasing amount of mutation data. When we have molecular sequence data which provide information about the branch length, the situation changes. Figure 3.4 shows the joint posterior distribution of rate and time for two example data sets. The posterior variances of time and rate are reduced when more molecular sequence data are available. The marginal posterior distribution and the prior of time t are plotted in Figure 3.5(a). Comparing the prior and the posterior distribution, information from the molecular sequence data pulls the mode of the time toward the true time $t = 1$. As indicated in Dos Reis and Yang[148], the marginal posterior distributions of time t are very close for $n = 100$ and $n = \infty$ (Figure 3.5(a)). In other words, the uncertainty of time is slightly reduced with the increase of sequence length when no mutation information is available.

However, Figure 3.4 and Figure 3.5(a) show that the posterior variances of time and rate are much reduced when mutation data are available. In addition, increasing the amount of mutation data can largely improve the inference of divergence time. Finite sequence data with infinite mutation data result in the smallest variances in time among all other combinations in Figure 3.5(a). The effect of changing the amount of mutation data in variance reduction is displayed in Figure 3.5(b). Compared with the case where infinite sequence data but no mutation data are used, having some sequence and mutation data results in more reduction in uncertainty in time. This suggests that spending money on obtaining longer molecular sequence data is not as cost efficient as obtaining some additional mutation data when the species analyzed comply with the molecular clock.

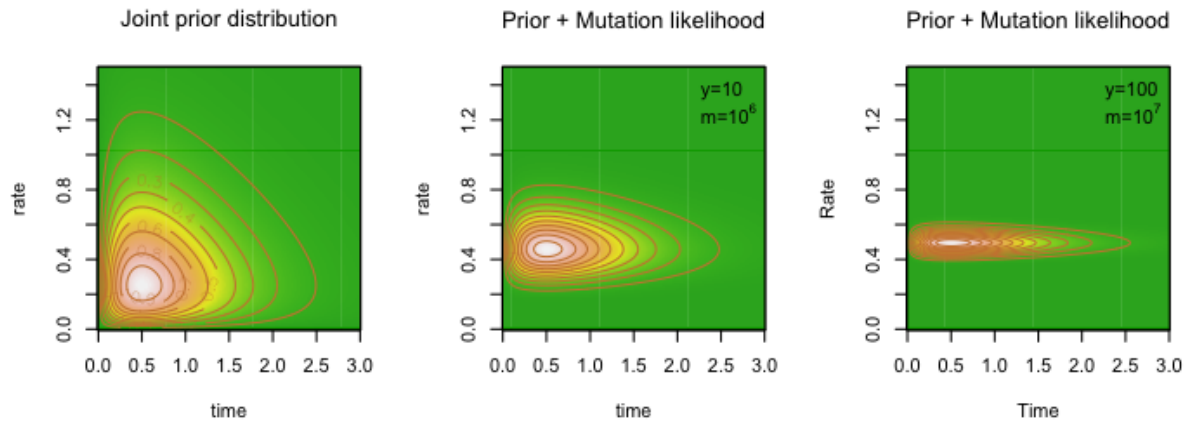


Figure 3.3: Joint prior distribution of the rate and the time and the effect of the mutation data on the prior under the strict clock. The mutation data consist of $y = 10$ or $y = 100$ mutations in a total of $m = 10^6$ or $m = 10^7$ bases assayed. The prior on the time remains diffuse regardless of the rate information provided by the mutation data. The prior on the rate changes dramatically when mutation data are available. Increasing the size of the mutation data from $m = 10^6$ to $m = 10^7$ can still narrow down the variance of the rate prior, but it has almost no influence on the time prior.

3.3.2 Autocorrelated Lognormal Relaxed Clock

Consider a case of 2 species with the rates at the "nodes" evolving under the autocorrelated lognormal relaxed clock . That is, the rate at the node drifts over time according to a geometric Brownian motion process. Because we will examine the effect when mutation data are associated with tips of the tree, we consider a special case where a geometric Brownian motion process is conditioned on the beginning and the end of a branch. In other words, we assume the logarithm of node rates drift according to a Brownian bridge starting at value $y_0 = \log(r_0)$ and stopping at $y_1' = \log(r_1')$, and $y_2' = \log(r_2')$, respectively, where r_0 is the rate of the root and r_1', r_2' are the rates of the two tips (Figure 3.6).

According to the formula provided by Guindon[147], we can calculate the expected value of the average rate along a branch under a geometric Brownian bridge given the rates of the root and the tip. Due to the stochasticity of a geometric Brownian motion process, the mean rate over a branch is a random variable. Therefore, we will assume the true mean rate of the branch is equal to the expected value of the mean rate. We use the product of this expected mean rate value and the true divergence time to be the true branch length. If the rate of the root and the tips are all 0.5 (i.e. $r_0 = r_1' = r_2' = 0.5$), the divergence time is $t = 1$, and the

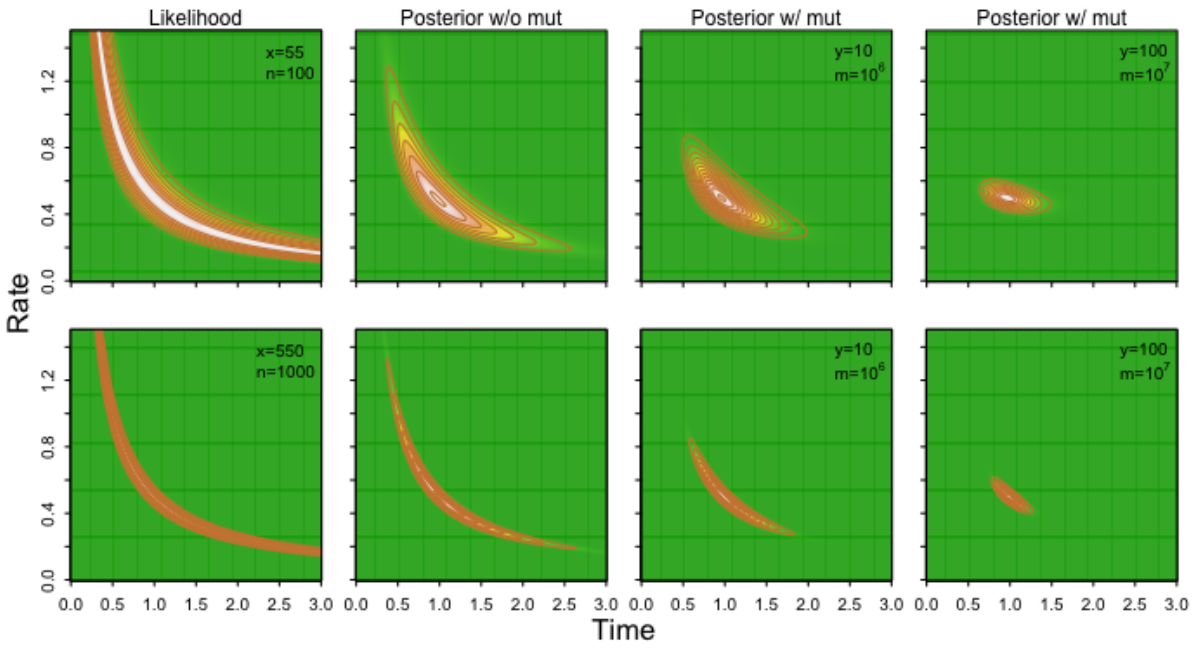


Figure 3.4: Likelihood and posterior distribution of the rate and the time for two datasets of a pairwise sequence alignment under the strict clock. The sequence data consist of x differences in an sites in the pairwise alignment. The figures of the first row are from a smaller sequence data set with $x = 55, n = 100$ and the figures of the second row are from a larger sequence data set with $x = 550, n = 1000$. Two sets of the mutation data are used as well, one with $m = 10^6$ and the other with $m = 10^7$ bases scanned.

autocorrelation parameter is $\nu = 0.25$, then the expected value of the average rate on each of the two branches is

$$E(\bar{r}_1) = \frac{1}{t}\exp(y_0)\int_0^t \exp\left(\frac{(y_1 - y_0)s}{t} + \frac{\nu(t-s)s}{2t}\right) ds$$
$$= 0.5\int_0^1 \exp\left(0.125 \times (1-s)s\right) ds = 0.51. \tag{3.23}$$

Therefore, the two branches both have branch lengths assumed to be $b = 2 \times 0.51 \times 1 = 1.02$.

Consider a gamma prior on the root rate $r_0 \sim G(2, 4)$ with mean 0.5 and variance 0.125, and a gamma prior on the time $t \sim G(2, 2)$ with mean 1 and variance 0.5. The rate prior selected is the autocorrelated lognormal relaxed clock formulated by Rannala and Yang[47], where the rate at the midpoint of a branch is used as a proxy for the average rate over the branch. Let r_1 and r_2 be the rates at the midpoints of the two branches (Figure 3.6), then the joint prior of r_1, t and r_1, r_0 can be seen in Figure 3.7. Note that r_1 and r_0 are highly positively correlated.

We again fit the data with the Jukes and Cantor model, where data are summarized with x differences in n sites in a pairwise alignment[49]. The likelihood of observing the sequence alignment given rate r_1, r_2 and time t is

$$\begin{aligned} L(r,t|n,x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \left(\frac{3}{4} - \frac{3}{4} e^{-4(r_1+r_2)t/3}\right)^x \left(\frac{1}{4} - \frac{3}{4} e^{-4(r_1+r_2)t/3}\right)^{n-x}. \end{aligned} \tag{3.24}$$

We will examine the effect of mutation data from one species on the inference of times by using two sets of mutation data on Species 1, each contains y mutations detected in m sites scanned. Similar to the strict clock case, $y = 10, m = 10^6$ and $y = 100, m = 10^7$, with generation time of 10 years are assumed. For simplification, we assume the rate at the tip is equivalent to the rate of the corresponding branch when doing the analysis. The rate information from the mutation data on tip 1 can reduce the uncertainty in branch rate r_1 (see Figure 3.7 for two different sets of mutation data).

Figures 3.8 and 3.9 show the joint posterior distributions of t and r_1, r_2, respectively, for two sequence data sets ($x = 55, n = 100$ and $x = 550, n = 1000$). Note that the joint posterior distributions of t, r_1 and t, r_2 are identically diffuse without mutation data. In contrast, the posterior variances of t and r_1 are reduced when mutation data are available for Species 1 as shown in Figure 3.8. However, the effect of mutation data is not as significant as the strict clock case. In addition, the effect of mutation data on reducing the uncertainty in rate r_2 is not as much as r_1 (Figure 3.9). In fact, as the amount of mutation data increases for Species 1, the uncertainty in r_2 remains almost the same. The rate information of Species 1 can influence the rate r_2 through reducing the uncertainty of root rate r_0. The influence of mutation data on the inference of r_2 depends on how much autocorrelation is present between rates and the times that separate these rates. If there is strong autocorrelation between rates (smaller ν) or if the time separates the two species is shorter (smaller t), then the rate information from Species 1 will have a greater influence on rate r_2.

Note that even if infinite sequence and mutation data from one species are available, the divergence time t depends on the free parameter r_2 and remains uncertain. Only if there are infinite mutation data on both branches can the divergence time t be known. The marginal posterior distribution of time t with different combination of sequence and mutation data can be seen in Figure 3.10(a).

3.4 Computer Simulations for Eight Species Case

The simulation design is aimed to examine the effect of mutation data with varying lengths of molecular sequence alignments under different circumstances: either alignments are generated under the strict clock or the autocorrelated relaxed clock, and whether the underlying phylogeny is balanced or not. To demonstrate the influence of the shape of the phylogeny, we used two different phylogenies of eight species, balanced and unbalanced trees, as shown in Figure 3.11. Molecular sequence data were generated using the EVOLVER program in the PAML package[12] and Jukes and Cantor model was assumed without rate heterogeneity among sites. For the strict clock case, the rate was assumed to be 1 nucleotide substitution per time unit, and our time units represent 100 million years (Myr). For the autocorrelated lognormal relaxed clock, the root rate was 1 nucleotide change per time unit and the rate variation among branches was modeled by a lognormal distribution with variance parameter $\nu = 0.1$. That is, the autocorrelated rates for branches described in Section 3.2.2 was used to sample the true rates among branches for each replicate. For each replicate, the unique rates sampled were multiplied by the true divergence times to generate branch lengths as the input for EVOLVER.

The mutation data consisted of the number of mutations that occurred per generation in a total 10^8 bases scanned for each species. We set the generation times for all species to be 20 years. The number of mutations that occurred for each species was generated from a Poisson distribution with mean $N \times \mu$, where N is the total number of bases scanned and μ is the mutation rate per site per generation. So if the true rate is r substitutions per site per 100 Myr, the mean of the Poisson distribution is $10^8 \times r \times \frac{1}{100 \times 10^6} \times 20 = 20r$. For the strict clock, mutation data were assigned only to Species A. For the autocorrelated relaxed clock, mutation data were placed either only on Species A or all species. Note that the rates on branches were sampled for each replicate for the autocorrelated relaxed clock case, so the rates were not the same for each replicate and thus the mutation data varied among replicates.

One hundred alignments were simulated with lengths $n = 10, 10^2, 10^3, 10^4$ and then analyzed with the program BEAST[14] to generate posterior samples of divergence times. The Jukes and Cantor model and the birth-death process with the default priors on the birth rate and the relative death rate were used in BEAST. The prior for the rate (root rate in the autocorrelated relaxed clock) was gamma with mean and standard deviation equal to 1 substitutions per site per 100 Myr. The root time prior was a gamma distribution with mean and standard deviation equal to 100 Myr and 20 Myr, respectively. And finally, the variance parameter in the autocorrelated relaxed clock, ν, had a gamma prior with mean and standard deviation equal to 0.001. Note that because we used one million years as one time unit in BEAST, we scaled the variance parameter due to time scale change. To reduce the computational burden, the tree topologies were fixed at all times. The average means and widths of 95% credible intervals for

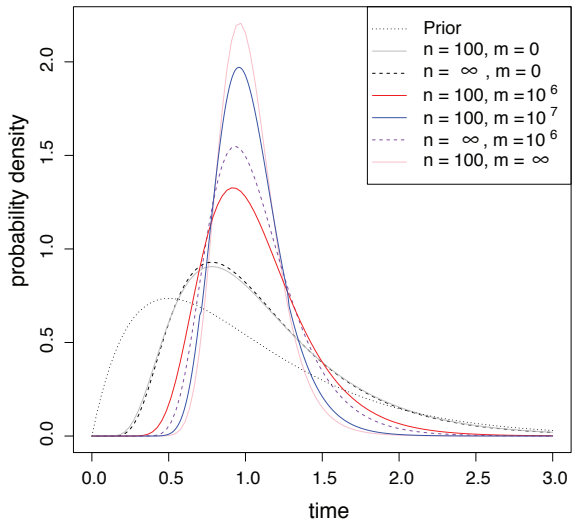
divergence times between Species A and Species B, Species A and Species D and the root time among 100 replicates were reported.

The results are shown in Table 3.1 - Table 3.4. We find that the tree shape was an issue. The inferred divergence times were more biased for the unbalanced tree than the balanced tree when the length of the aligned sequences was short ($n = 10$ or 100), regardless of whether the strict clock or the autocorrelated relaxed clock was assumed. For longer alignments ($n = 1000$ or 10000), the effect of the tree shape on the estimated divergence times was minor for the strict clock case. But for the autocorrelated relaxed clock case, the estimated divergence times of Species A and Species B, and Species A and Species D were more biased when the true phylogeny was an unbalanced tree than a balanced tree with longer alignments and no mutation data. Since the widths of the 95% credible intervals for the divergence times were proportional to the magnitude of estimated divergence times, they were narrower for the unbalanced tree case than the balanced tree case. These results could be due to the method for sampling rates under the autocorrelated relaxed clock. The terminal branch rate of Species A for the balanced tree was sampled conditioned on its ancestral branch rate, which is itself conditioned on its ancestral branch rate. This kind of conditioning repeated 3 times to sample the terminal branch rate of Species A for the balanced tree, but repeated 7 times for the unbalanced tree. The terminal branch rate of Species A for the unbalanced tree could vary a lot between different simulated samples than for the balanced case, which led to more biases in divergence times of Species A and Species B, and Species A and Species D, especially when no mutation data are present.

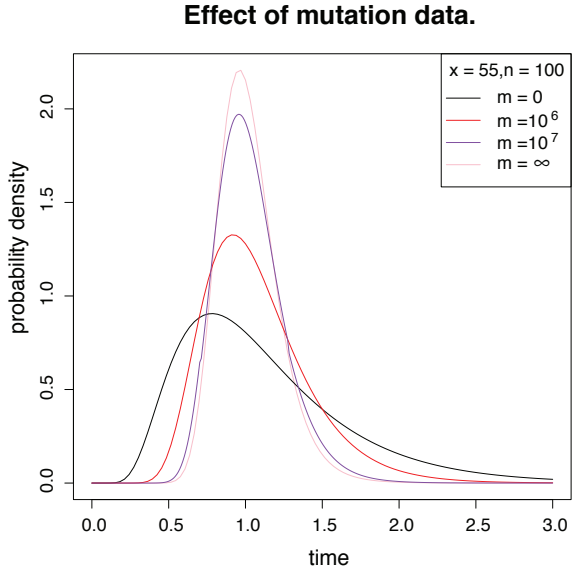
Second, the influence of the mutation data was noticeable. The underestimation of the root time was reduced when mutation data were available, under both the strict and the autocorrelated relaxed clock. Moreover, the widths of the 95% credible intervals for all the divergence times were substantially shorten with mutation data. The widths of the 95% credible intervals for divergence times and the estimated divergence times were in direct proportion, and so was the amount of shrinkage of the widths because of the mutation data.

Third, the mutation data and the tree shape affected the estimated divergence times differently with the autocorrelated clock than with the strict clock. The effect of the mutation data on the widths of the 95% credible intervals was similar between the balanced and the unbalanced tree under the strict clock. In contrast, the influence of the mutation data was distinct under the autocorrelated clock. The estimated divergence times and the widths of the 95% credible intervals of Species A and Species B were alike between the balanced and the unbalanced tree if mutation data were available for Species A. However, the 95% credible interval of the divergence time of Species A and Species B was narrower under the unbalanced tree than under the balanced tree, although the mutation data were available for both Species A and Species B. In addition, the 95% credible interval of the divergence time of Species A and Species D was narrower under the unbalanced tree than under the balanced tree.

Finally, if mutation information were available for all the extant species, the widths of the 95% credible intervals for the divergence times reported were significantly reduced compared to the case with only mutation information for one species.



(a) The marginal posterior of the time with different combination of sequence and mutation datasets.



(b) The effect of the mutation data on the marginal distribution of the time.

Figure 3.5: The marginal posterior distribution of the time under the strict clock. Without mutation information, increasing the sequence length from $n = 100$ to infinity has little effect on the posterior distribution of the time. The effect of the mutation data can be seen from the decrease in uncertainty in the marginal posterior distribution of the time.

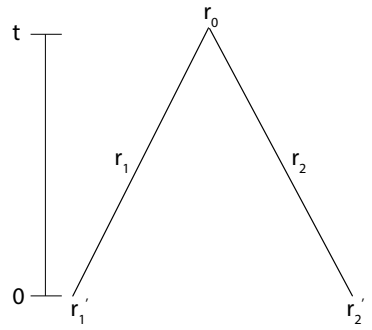


Figure 3.6: The rooted tree of 2 species. The root rate is r_0. r_1' and r_2' are the rates at the tips. r_1 and r_2 are the rates at the midpoint of the branches.

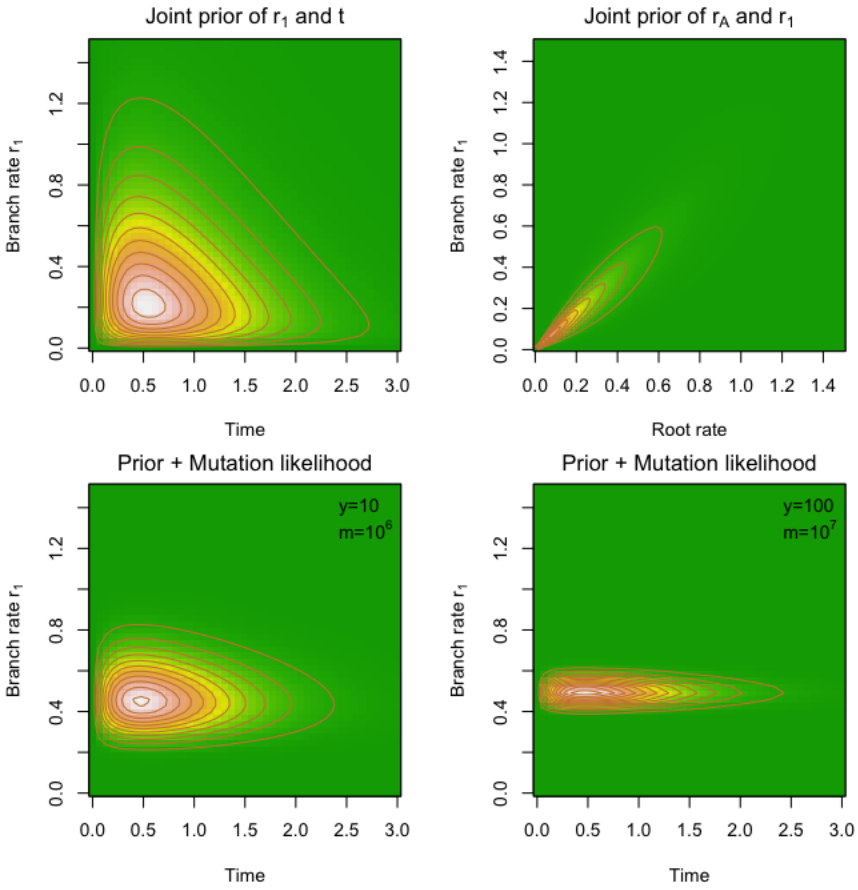


Figure 3.7: The joint prior and the effect of mutation data on the prior under the autocorrelated relaxed clock model. The ancestral rate and the rate at the midpoint of the child branch are correlated in the prior due to the autocorrelated relaxed clock model assumption. Two sets of mutation data are used, one with $y = 10$ mutations in a total of $m = 10^6$ bases scanned, and the other with $y = 100$ mutations in a total of $m = 10^7$ bases scanned. Similar to the strict clock case, increasing the size of the mutation data from $m = 10^6$ to $m = 10^7$ can narrow down the variance of the rate prior, but it has almost no influence on the time prior.

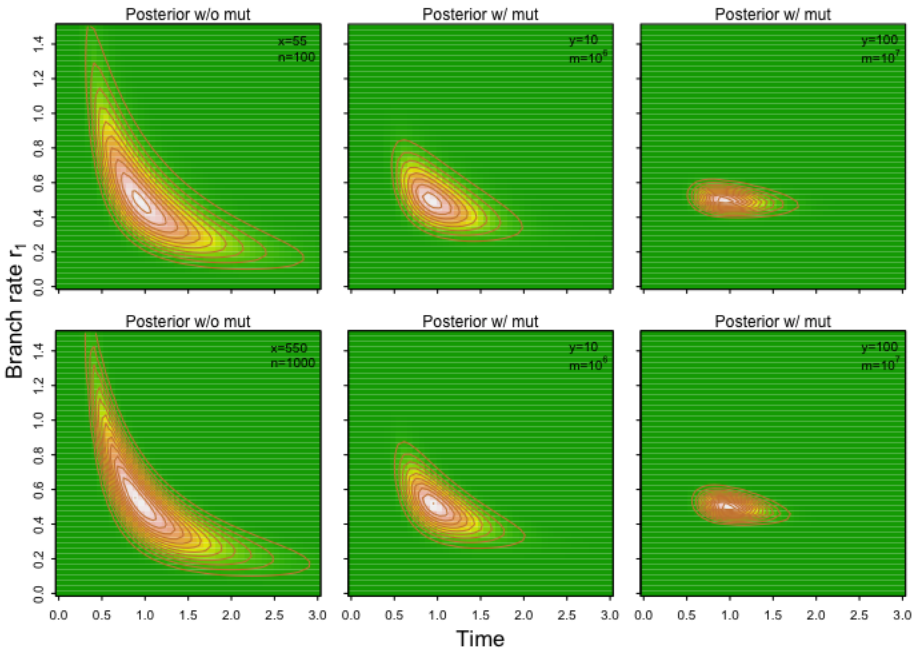


Figure 3.8: The posterior distribution of the branch rate r_1 and the time for two datasets of a pairwise sequence alignment under the autocorrelated relaxed clock model. The figures on the first row are from $x = 55, n = 100$ and the figures on the second row are from $x = 550, n = 1000$. The mutation data are placed on branch 1. There are two sets of mutation data, one with $y = 10$ observed mutations in $m = 10^6$ bases scanned, and the other with $y = 100$ mutations in $m = 10^7$ bases scanned. The joint posterior distribution of the branch rate r_1 and the time with the mutation data are shown in the second and the third column.

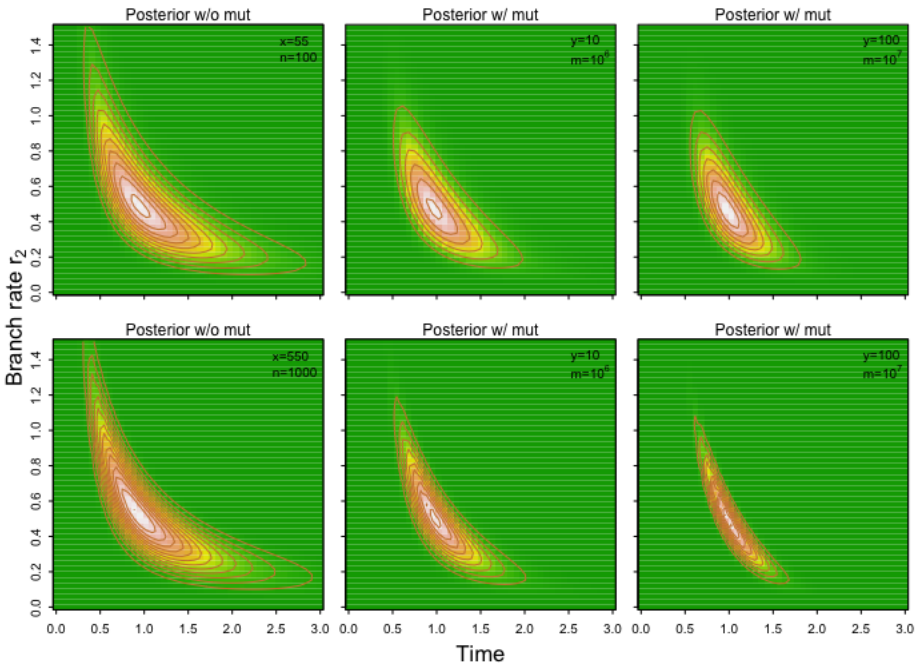
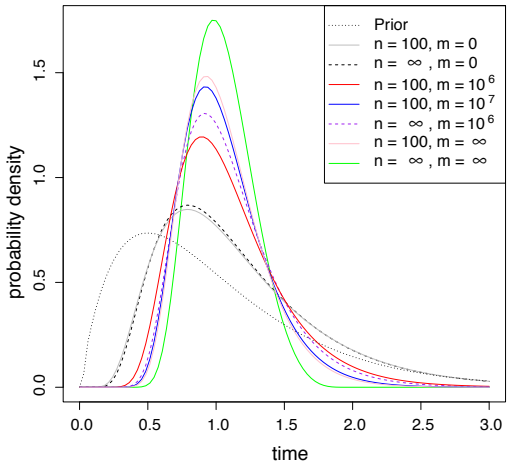
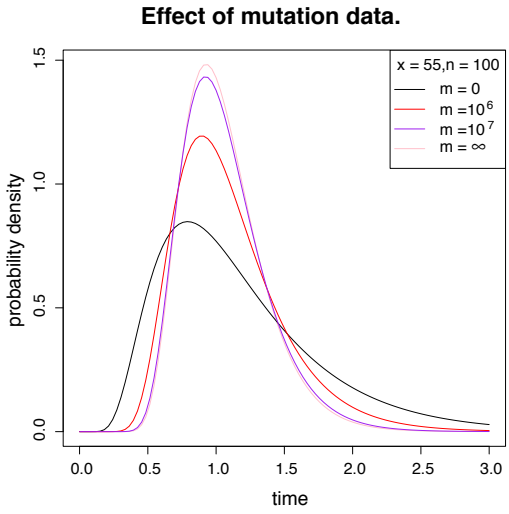


Figure 3.9: The posterior distribution of the branch rate r_2 and the time for two datasets of a pairwise sequence alignment under the autocorrelated relaxed clock model. The figures on the first row are from $x = 55, n = 100$ and the figures on the second row are from $x = 550, n = 1000$. The mutation data are placed on branch 1. There are two sets of mutation data, one with $y = 10$ observed mutations in $m = 10^6$ bases scanned, and the other with $y = 100$ mutations in $m = 10^7$ bases scanned. The joint posterior distribution of the branch rate r_2 and the time with the mutation data are shown in the second and the third column.

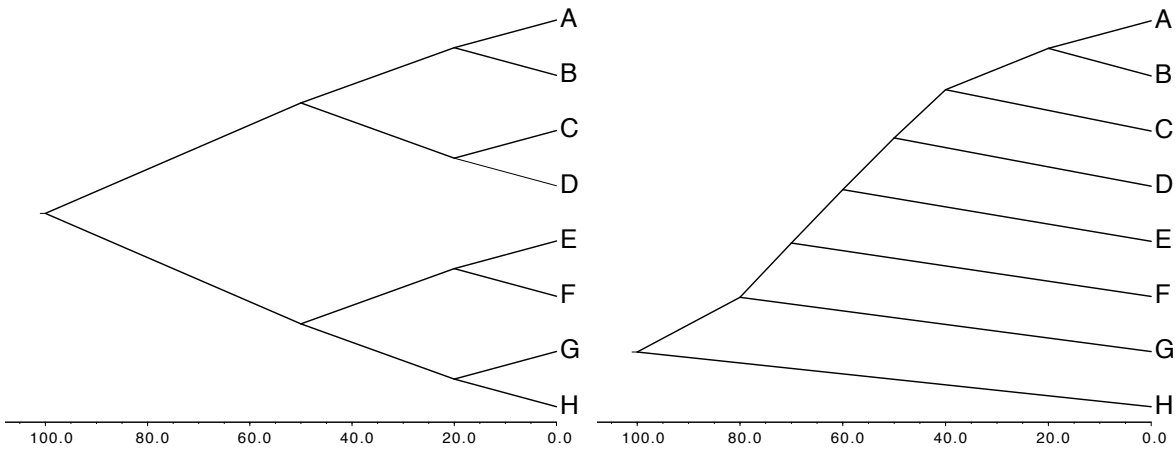


(a) The marginal posterior of the time with different combination of sequence and mutation datasets.



(b) The effect of the mutation data on the marginal distribution of the time.

Figure 3.10: The marginal posterior distribution of the time under the autocorrelated relaxed clock model. Without mutation information, increasing the sequence length from $n = 100$ to infinity has little effect on the posterior distribution of the time. On the other hand, increasing the amount of the mutation data on branch 1 does decrease the uncertainty in time. However, it can also be seen from this figure that infinite sequence data as well as infinite mutation data on branch 1 does not make the uncertainty in time shrink to zero.



(a) Balanced tree

(b) Unbalanced tree

Figure 3.11: Two tree topologies of 8 species used in the simulations. The time on the scale is in million years.

Table 3.1: Divergence times for a balanced tree and a strict clock. The true phylogeny is in Figure 3.11(a). The means and the widths of the 95% posterior CIs of divergence times with varying alignment lengths are reported. The "No Mut" rows display the results without mutation information. The "Mut" rows display the results with the mutation data generated from a Poisson distribution.

		Time A & B (t=20)				Time A & D (t=50)				Root time (t=100)			
	n	10	100	1000	10000	10	100	1000	10000	10	100	1000	10000
No Mut	Mean	15.72	18.56	18.68	18.62	35.52	46.92	46.41	46.93	91.69	92.01	91.59	91.73
	95% C.I. width	35.46	23.03	15.62	14.26	67.16	53.27	37.8	35.9	72.66	71.74	70.87	69.84
Mut	Mean	16.06	20.25	19.42	19.83	37.08	51.27	48.31	50	91.65	96.37	95.39	97.77
	95% C.I. width	33.17	19.42	12.08	11.22	59.73	42.46	28.85	28.24	66.71	60.22	54.18	55.08

Table 3.2: Divergence times for an unbalanced tree and a strict clock. The true phylogeny is in Figure 3.11(b). The means and the widths of the 95% posterior CIs of divergence times with varying alignment lengths are reported. The "No Mut" rows display the results without mutation information. The "Mut" rows display the results with the mutation data generated from a Poisson distribution.

		Time A & B (t=20)				Time A & D (t=50)				Root time (t=100)			
	n	10	100	1000	10000	10	100	1000	10000	10	100	1000	10000
No Mut	Mean	7.81	26.38	18.01	18	22.23	47.79	45.56	45	94.02	91.92	91.58	91.44
	95% C.I. width	17.85	30.31	15.31	14.22	40.67	49.49	38.11	35.44	73.45	71.9	72.45	71.68
Mut	Mean	9.6	27.89	18.64	19.26	27.54	50.45	47.2	48.16	96.92	94.95	94.74	97.92
	95% C.I. width	19.39	24.84	11.62	10.85	38.35	36.78	28.46	27.06	69.1	58.91	54.43	54.63

Table 3.3: Divergence times for a balanced tree and an autocorrelated lognormal relaxed clock. The true phylogeny is in Figure 3.11(a). The means and the widths of the 95% posterior CIs of divergence times with varying alignment lengths are reported. The "No Mut" rows display the results without mutation information. The "One Mut" rows display the results with the mutation data for Species A, which were generated from a Poisson distribution. The "All Mut" rows are for the results with mutation data available for all the extant species.

		Time A & B (t=20)				Time A & D (t=50)				Root time (t=100)			
	n	10	100	1000	10000	10	100	1000	10000	10	100	1000	10000
No Mut	Mean	21.57	18.94	19.12	19.15	49.44	45.52	47.46	47.32	93.03	92.69	92.29	93.11
	95% C.I. width	45.17	27.18	21.45	20.09	77.31	56.7	47.93	46.71	72.91	72.29	71.92	71.63
One Mut	Mean	22.14	20.31	20.06	19.71	50.93	49.14	49.99	48.85	93.5	95.91	96.04	96.34
	95% C.I. width	41.4	21.22	15.31	13.44	69.02	45.19	37.04	33.29	68.85	64.71	61.82	60.23
All Mut	Mean	22.8	20.83	20.12	20.22	52.74	50.52	50.51	50.11	94.16	97.66	98.88	98.63
	95% C.I. width	41.28	18.91	11.35	9.56	66.64	36.03	23.5	21.58	65.61	51.96	37.93	34.53

Table 3.4: Divergence times for an unbalanced tree and an autocorrelated lognormal relaxed clock. The true phylogeny is in Figure 3.11(b). The means and the widths of the 95% posterior CIs of divergence times with varying alignment lengths are reported. The "No Mut" rows display the results without mutation information. The "One Mut" rows display the results with the mutation data for Species A, which were generated from a Poisson distribution. The "All Mut" rows are for the results with mutation data available for all the extant species.

		Time A & B (t=20)				Time A & D (t=50)				Root time (t=100)			
	n	10	100	1000	10000	10	100	1000	10000	10	100	1000	10000
No Mut	Mean	10.76	14.81	16.35	16.59	29.23	36.57	41.34	41.84	94.95	92.39	91.66	91.32
	95% C.I. width	23.48	21.68	18.31	17.5	49.23	46.61	43.3	42.05	73.52	71.96	71.44	71.13
One Mut	Mean	14.99	18	19.02	18.97	38.95	44.29	47.98	47.76	101.21	99.09	99.23	98.8
	95% C.I. width	27.26	18.09	13.51	12.54	48.1	37.48	33.25	31.14	71.44	65.84	62.86	61.68
All Mut	Mean	15.44	19.61	19.89	19.65	40.77	48.4	50.1	49.67	102.14	103.01	100.57	101.04
	95% C.I. width	27.18	16.71	9.35	7.55	45.79	28.26	19.51	16.84	67.94	52.96	40.78	38.48

3.5 Discussion and Conclusions

Mutation data can provide information about the rates of evolution, and therefore reduce the amount of uncertainty about rates in divergence time estimation. However, less uncertainty about rates is not guaranteed to yield comparatively less uncertainty about times. The uncertainty in times depends on prior information, the rate information from the mutation data, and the likelihood of the interspecific sequence data (Figure 3.1).

Even if branch lengths could be perfectly estimated from infinitely long sequences that diverged via a strict clock and a known model of nucleotide substitution, variances in the posterior distribution of rates and times will be nonzero unless there is additional fossil or mutation information available for separating rates and times. When infinitely long sequences and a strict clock are coupled with infinitely much mutation data, evolutionary rates and divergence times can be identified without error.

On the other hand, if a strict clock is not valid and relaxed clocks are used, then having an infinite amount of mutation data for one species will only help us know the rate of the tip associated with species. The degree to which the mutation data is informative for separating rates and times depends on the level of rate autocorrelation among branches. With rate autocorrelation, mutation data on one tip is relatively informative about rate for adjacent branches that separate closely related species. It becomes less informative about rates for branches that are far from the tips where mutation data are available.

If mutation data are available for every extant species, the divergence times of Species A and B, C and D, E and F, G and H on the balanced tree (Figure 3.2(b)) can have less associated uncertainty by having double the amount of mutation information from both species. The uncertainty of the deeper divergence times for other combinations of species on the balanced tree relies on the strength of the rate signals on the tips and the level of rate autocorrelation among branches. On the other hand, mutation data from extant species (e.g. Species H) on the unbalanced tree (Figure 3.2(c)) can provide rate information for branches with deeper divergence times (e.g. root-to-H branch). The shape of the underlying phylogeny could affect the rate autocorrelation structure and how mutation rate information on tips of the tree acts on divergence time estimation. Its effect is complicated and needs to be investigated case by case.

Last but not least, mutation rates vary among chromosomes[149] and genomic regions because of differences in base composition and recombination rates[150]. There are some mutational hotspots within the genome (e.g. CpG sites[23, 109]), and mutation rates are generally higher in synonymous sites in the coding regions than in noncoding regions[22]. Mutation rates also vary among individuals[102, 109] due to environment[151, 152], parental age[109], sex[153, 151, 154, 21] and many other factors. However, mutation data are usually collected

from whole genome analyses. Experimentally derived mutation rates are thus averages over different genomic regions and over different individuals, and they may not be representative of the region used to estimate divergence times. Future research about mutation mechanism can allow incorporation of the uncertainty in the mutation rates and improve divergence time estimation.

Chapter 4

Relaxing the molecular clock to different degrees for different substitution types

4.1 Abstract

Rates of molecular evolution can vary over time. Diverse statistical techniques for divergence time estimation have been developed to accommodate this variation. These typically require that all sequence (or codon) positions at a locus change independently of one another. They also generally assume that the rates of different types of nucleotide substitutions vary across a phylogeny in the same way. This permits divergence time estimation procedures to employ an instantaneous rate matrix with relative rates that do not differ among branches. However, previous studies have suggested that some substitution types (e.g., CpG to TpG changes in mammals) are more clock-like than others. As has been noted by others, this is biologically plausible given the mutational mechanism of CpG to TpG changes. Via stochastic mapping of sequence histories from context-independent substitution models, our approach allows for context-dependent nucleotide substitutions to change their relative rates over time. We apply our approach to the analysis of an 0.15 Megabase region from 8 primates. In accord with previous findings, we find comparatively little rate variation over time for CpG to TpG substitutions but we find more for other substitution types. We conclude by discussing the limitations and prospects of our approach.

4.2 Introduction

Molecular sequence data provide information about the amount of evolution, or the branch lengths, between species, but they do not suffice for disentangling rates and times. If, however, sequences evolve at a constant rate, then inferring divergence times with fossil and molecular data becomes relatively straightforward because separation of rates and times on one branch of a phylogenetic tree allows separation on other branches as well. Even though the chronologically constant rate of a molecular clock has been widely assumed in phylogenetic studies, mutations do not arise uniformly throughout the genome and the clock ticks at different rates in different genomic regions. In addition, different proteins may change according to their own molecular clocks because of natural selection. Moreover, different parts within the same protein can even evolve at distinct rates, depending on the strength of the selective constraint (either functional or structural) at that region.

In addition to rate variation among different genomic regions, the molecular clock does not tick at the same rate for all phylogenetic lineages. The rate at which sequences change depends on a large number of factors (e.g., population size, generation time, mutation rate). These factors vary among lineages. When substitution rates change over time, divergence time estimation becomes more challenging because separation of rates and times on one part of a phylogeny does not guarantee that rates and times can be well-separated on other parts of the tree.

A wide variety of relaxed clock methods have been proposed and studied for facilitating this separation[41, 9, 42, 43, 10, 44, 45, 46, 47]. One promising direction is to identify factors that cause rate variation over time or that at least covary with this rate variation. For example, Lartillot and Delsuc[155] have considered how life history traits such as body size might covary with nucleotide substitution rates. By combining life history trait information of extant species with the sequence data from these species, Lartillot and Delsuc are able to partially separate evolutionary rates and times by exploiting covariation across a phylogeny between body size and evolutionary rates.

Other potential ways to partially or completely separate rates and times are motivated by other factors that affect substitution rates. These include the possibility that natural selection induces a correlation between substitution rate and effective population size[156] and the ability to better understand substitution rates by incorporating mutation data (see Chapters 2 and 3). Although diverse treatments of substitution rate variation over time are available, a typical assumption is that the relative rates of different substitution types vary among branches on a phylogenetic tree in relatively simple ways. For example, the rates of all substitution types at one instant in time might be exactly some multiplicative factor of the corresponding rates at some other instant. Alternatively, effective population size might vary among branches of a

tree so that mutation-selection balance approaches inspired by population genetics can assist in the determination of how rates of different substitution types vary among branches (e.g., see Yang and Nielsen[119] and Rodrigue et al.[120]).

Beyond the effects of generation time and natural selection, there are compelling biological reasons to believe that additional flexibility in substitution rate variation over time is warranted. Even for a neutrally evolving sequence, variation across the phylogeny in generation length might be insufficient to completely explain variation across the phylogeny in substitution rates. Although mutation rate equals substitution rate with neutral evolution[30, 31], different mutational mechanisms underlie different types of point mutations. If mutation occurred only in meiosis, then mutation rate variation over time might be exclusively attributable to changes in generation length. But, this is not the situation. Some mutational mechanisms seem prone to yielding almost clock-like substitution behavior because they can operate nearly at any time during a life cycle whereas other mutational mechanisms are meiotic and therefore prone to affecting a sequence site only once per generation. In addition, the relative importance of mutational mechanisms can change across a phylogeny because the genetic and environmental components affecting them can change.

Despite the fact that the molecular clock seldom holds in empirical analyses, it has been suggested that CpG dinucleotides may evolve in a comparatively clock-like fashion in mammals and primates[157, 158, 159]. Since DNA methylation could occur in the early development stage of germ-line cells or embryonic cells[23], mutations associated with DNA methylation can accumulate throughout the lifespan of germ-line cells[158]. Hwang and Green implemented a pioneering Bayesian Markov chain Monte Carlo method that allows substitution rates to depend on neighboring nucleotides. They observed that CpG transitions have a relatively clock-like behavior in mammals[157]. Kim et al. noted that divergence times between the human and chimpanzee species pair and the macaque and baboon species pair are thought to be similar, but the human-chimpanzee pair has longer generation times. They pointed out that macaque-baboon pair had accumulated significantly more transitions than the other pair in non-CpG sites, probably by virtue of the generation-time effect. However, inferred amounts of transition substitutions were similar between pairs in CpG sites[158]. As emphasized by the authors of these studies, methylation-origin mutations at CpG sites may be dependent on chronological time rather than on time measured in generations. In other words, CpG sites might serve as the basis of a relatively accurate molecular clock even for a phylogeny relating lineages with diverse generation times.

Inspired by these studies on CpG substitution and also by the realization that there are a variety of other reasons why substitution rates may vary over time differently for different substitution types, we aim to investigate the clock-like natures of different kinds of (possibly context-dependent) nucleotide substitutions and we aim to simultaneously improve divergence

time estimation.

4.3 Materials and Methods

4.3.1 Augmented Data Likelihood

The observed data are the homologous sequence alignment $\mathbf{D}$ of length N from S species. The species are related through a rooted phylogenetic tree via a tree topology that is assumed known. We will use the index i for sites and index j for nodes throughout the text. The root has index 0, tips have indices $j = 1, ..., S$, and internal nodes have indices $j = S+1, ..., 2S-2$. A branch will have the same index as the node at its end.

Consider a specific branch j on a tree and a specific site i in a DNA sequence. Assume the full substitution history of the DNA sequence on branch j is observed. The substitution rate per site on branch j from site context a to nucleotide b will be μ_{abj}. Here, "context" represents both the state of the site being followed and (possibly) states at other sites. The vector $\boldsymbol{\mu}_j$ will represent the entire collection of μ_{abj} for all possible contexts a and nucleotides b on branch j.

We are interested in the product of μ_{abj} and time for each branch, and we refer this product the "substitution length" of branch j from context a to nucleotide b. Although the product can be estimated from sequence data, rates and the branch time are confounded. We will set the branch time at 1 for all branches. This means that the substitution length for change from a to b is known on branch j if μ_{abj} is known. Let ϕ_{aij} ($\phi_{aij} \in [0, 1]$) be the proportion of time site i has context a on branch j with $\boldsymbol{\phi}_{ij}$ being a vector representing values of ϕ_{aij} for all a, and let n_{abij} be the number of changes from context a to nucleotide b at site i on branch j and with $\mathbf{n}_{ij}$ being the vector of n_{abij} for all a and b. If the nucleotide substitution process is represented as a continuous time Markov chain, then the likelihood for the observed history of site i on branch j is

$$p(\mathbf{n}_{ij}, \boldsymbol{\phi}_{ij} | \boldsymbol{\mu}_j, s_{ij0}) = \prod_a \prod_b \mu_{abj}^{n_{abij}} \exp\{-\phi_{aij}\mu_{abj}\}, \tag{4.1}$$

where s_{ij0} is the initial state of site i on branch j. The likelihood without conditioning on the initial state s_{ij0} would be

$$p(\mathbf{n}_{ij}, \boldsymbol{\phi}_{ij} | \boldsymbol{\mu}_j, s_{ij0}) p(s_{ij0} | \boldsymbol{\mu}_j) = p(\mathbf{n}_{ij}, \boldsymbol{\phi}_{ij}, s_{ij0} | \boldsymbol{\mu}_j).$$

One simplification that is often made when modeling molecular evolution is to assume stationarity of the substitution process. The stationarity assumption means that the initial state s_{ij0} has some information pertaining to the rate parameters. Rather than assume stationarity, we treat $p(s_{ij0}|\boldsymbol{\mu}_j)$ as free parameter. This means that Equation 4.1 summarizes the information

about $\boldsymbol{\mu}_j$ in our likelihood.

Suppose the complete substitution history of every site is available, then the likelihood for the observed history of all sites on branch j is then

$$p(\mathbf{n}_j, \boldsymbol{\phi}_j | \boldsymbol{\mu}_j, \mathbf{s}_{j0}) = \prod_i \prod_a \prod_b \mu_{abj}^{n_{abij}} \exp\{-\phi_{aij}\mu_{abj}\} \tag{4.2}$$

$$= \prod_a \prod_b \mu_{abj}^{\sum_i n_{abij}} \exp\{-\mu_{abj} \sum_i \phi_{aij}\}, \tag{4.3}$$

where $\mathbf{n}_j$, $\boldsymbol{\phi}_j$ and $\mathbf{s}_{j0}$ are respectively vectors of $\mathbf{n}_{ij}$, $\boldsymbol{\phi}_{ij}$ and s_{ij0}. The log-likelihood is then

$$\log p(\mathbf{n}_j, \boldsymbol{\phi}_j | \boldsymbol{\mu}_j, \mathbf{s}_{j0}) = \sum_i \sum_a \sum_b n_{abij} \log\{\mu_{ab}\} - \phi_{aij}\mu_{ab} \tag{4.4}$$

$$= \sum_a \sum_b N_{abj} \log\{\mu_{ab}\} - \Phi_{aj}\mu_{ab}, \tag{4.5}$$

where $N_{abj} = \sum_i n_{abij}$ and $\Phi_{aj} = \sum_i \phi_{aij}$. By setting the first derivative to zero and solving, we have the maximum likelihood estimates

$$\hat{\mu}_{abj} = \frac{N_{abj}}{\Phi_{aj}}.$$

In other words, the sufficient statistics of μ_{abj} on branch j consists of the total time spent in context a on branch j, Φ_{aj}, and the number of changes from context a to nucleotide b on branch j, N_{abj}.

We can consider $\boldsymbol{\mu}_{ab}$ as a vector of substitution lengths of changes from context a to nucleotide b on the phylogenetic tree.

$$\boldsymbol{\mu}_{ab} = \begin{pmatrix} \mu_{ab1} \\ \vdots \\ \mu_{ab(2S-2)} \end{pmatrix} = (\mu_{ab1}, ..., \mu_{ab(2S-2)})^T. \tag{4.6}$$

Instead of different genes having different branch lengths, now each type of change has its own "substitution length." The vectors with these substitution lengths for different branches are represented by $\boldsymbol{\mu}_{ab}$ for context a and nucleotide b.

Let $\mathbf{M} = \{M_{ij}\}_{1 \le i \le N, 1 \le j \le 2S-2}$ represent the substitution history (mutation mapping) that generates the molecular sequence data $\mathbf{D}$ such that M_{ij} is the substitution history at site i on branch j. The substitution history M_{ij} can be summarized by the sufficient statistics for μ_{abj}: n_{abij} and ϕ_{aij}.

$$M_{ij} = (n_{abij}, \phi_{aij}), i = 1, 2, ..., N; j = 1, 2, ..., 2S-2. \tag{4.7}$$

With a context-independent model of sequence change, there are 4 possible states (A, T, C, G) and $4 \times 3 = 12$ possible types of single nucleotide substitutions (e.g. A $\rightarrow$ T, A $\rightarrow$ C, etc.) if context is not considered. If strand-symmetry is assumed so that types have the same rate as their complement (e.g., A $\rightarrow$ T and T $\rightarrow$ A have an identical rate), context-independence yields 6 types of single nucleotide substitutions: 4 transversions and 2 transitions. These 6 types of single nucleotide substitutions are listed in Table 4.1. On the other hand, there are 9 types of changes (pooling together complementary changes) and 3 states (contexts) when classifying a site according to whether it or its complement might have a methylated C in a CpG dinucleotide. These 9 types of changes listed in Table 4.2 include 4 transversions and 2 transitions for non-CpG sites, and 2 transversions and 1 transition for CpG sites. For instance, the substitution length of CpG transitions (Type 9 in Table 4.2) of a branch can be inferred by the number of CpG transitions divided by the proportion of time CpG sites were located on that branch if the substitution history is fully observed. Similarly, the substitution length of non-CpG G $\rightarrow$ C and C $\rightarrow$ G substitutions (Type 1 in Table 4.2) is computed as the number of G $\rightarrow$ C and C $\rightarrow$ G substitutions in non-CpG sites divided by the proportion of time that non-CpG C or G sites exist on that branch. Note that the way we define substitution lengths can be applied to other types of context-dependent substitutions not discussed here. It is possible to derive the substitution lengths for contexts considering any combinations of 5' and 3' neighboring nucleotides.

Table 4.1: The 6 single nucleotide substitution types. Complementary substitutions are considered as the same type.

Type	Substitutions
1	G $\rightarrow$ C & C $\rightarrow$ G
2	G $\rightarrow$ T & C $\rightarrow$ A
3	T $\rightarrow$ A & A $\rightarrow$ T
4	T $\rightarrow$ G & A $\rightarrow$ C
5	G $\rightarrow$ A & C $\rightarrow$ T
6	A $\rightarrow$ G & T $\rightarrow$ C

4.3.2 Sampling Substitution Histories

The 'substitution lengths' of different types of substitutions provide a basis for understanding the rates of each kind of change. This quantity is hard to estimate directly for the molecu-

Table 4.2: The 9 context-dependent CpG substitution types. Complementary substitutions are considered as the same type.

Type	Substitutions
1	Non-CpG G $\rightarrow$ C & C $\rightarrow$ G
2	Non-CpG G $\rightarrow$ T & C $\rightarrow$ A
3	Non-CpG T $\rightarrow$ A & A $\rightarrow$ T
4	Non-CpG T $\rightarrow$ G & A $\rightarrow$ C
5	Non-CpG G $\rightarrow$ A & C $\rightarrow$ T
6	Non-CpG A $\rightarrow$ G & T $\rightarrow$ C
7	CpG G $\rightarrow$ C & C $\rightarrow$ G
8	CpG G $\rightarrow$ T & C $\rightarrow$ A
9	CpG G $\rightarrow$ A & C $\rightarrow$ T

lar sequence alignment, but the estimation becomes rather straightforward if the substitution histories (mappings) can be fully observed. We therefore employ data augmentation to help estimate the 'substitution lengths' of each kind of change. That is, the augmented data (unobserved substitution histories) are constructed from the observed sequence alignment.

Diverse strategies have been developed for sampling substitution histories $\mathbf{M}$ conditional upon the sequence alignment $\mathbf{D}$ and a vector of parameters Ω that represents the tree topology and the parameters of the substitution process. Sampling histories $\mathbf{M}$ from the distribution $Pr_{\Omega}(\mathbf{M}|\mathbf{D})$ is categorized as endpoint-conditioned sampling because the sequence data $\mathbf{D}$ are observed at the endpoints (tips) of the tree. Endpoint-conditioned sampling strategies for molecular sequence data have been evaluated by Hobolth and Stone[160] and reviewed by Hobolth and Thorne[161]. We will not review methods for endpoint-conditioned sampling here.

The probability of a mapping conditioned on the molecular sequence alignment can be written as

$$Pr_{\Omega}(\mathbf{M}|\mathbf{D}) = Pr(\mathbf{M}|\mathbf{D},\Omega) = \frac{Pr(\mathbf{M},\mathbf{D},\Omega)}{Pr(\mathbf{D},\Omega))}. \tag{4.8}$$

However, the true values of the parameters in Ω are usually unknown. A mapping sampled from the marginal distribution $Pr(\mathbf{M}|\mathbf{D})$ by integrating out all possible parameter values in Ω has density

$$Pr(\mathbf{M}|\mathbf{D}) = \int_{\Omega} Pr(\mathbf{M}|\mathbf{D},\Omega)Pr(\Omega|\mathbf{D})d\Omega. \tag{4.9}$$

Lartillot proposed a sampling algorithm that includes data augmentation and conjugate Gibbs sampling to obtain samples from the joint posterior probability distribution $Pr(\mathbf{M},\Omega|\mathbf{D})$[162] and implemented it in the PhyloBayes software[163]. The algorithm proceeds in two alternating steps: first, drawing a substitution history conditional on the parameters in Ω similar to

Nielsen's algorithm[164], and second, resample the parameters in Ω by a Gibbs sampler conditional on the substitution history. A Gibbs sequence is generated after repeating this process many times, where a subset of samples of Ω and substitution histories are taken as draws from the full joint posterior distribution of all parameters. The Monte Carlo estimate of any moments (e.g. mean) for the marginal distribution of the substitution histories can be directly computed from the realizations of the Gibbs sequence.

We use the PhyloBayes software with an independent-site time-reversible model to generate the substitution histories. We then use these sampled histories to make inferences about parameters in a richer context-dependent substitution model. We do this by replying upon the assumption that the distribution of substitution histories is robust to substitution model specification. In other words, we assume that the endpoint-conditioned sampled by PhyloBayes for our data set can be treated as an endpoint-conditioned sample according to our substitution model of interest. For data sets with long branches that generate high probabilities of multiple changes per site or high probabilities of changes at consecutive sites, this assumption will be problematic. For data sets with short branches and little sequence divergence, the assumption should be more appropriate.

4.3.3 Substitution Length Estimation

The PhyloBayes[163] software implements a Markov chain Monte Carlo approach for sampling endpoint-conditioned substitution histories $\mathbf{M}$ and parameters $\Omega = (\mathbf{l}, \mathbf{r}, \boldsymbol{\rho}, \boldsymbol{\pi})$ from the joint posterior distribution $Pr(\mathbf{M}, \Omega|\mathbf{D})$. By only keeping a set of widely spaced realizations of mappings and parameters, we can generate C approximately independent and identically distributed samples of $\mathbf{M}$ and Ω from $Pr(\mathbf{M}, \Omega|\mathbf{D})$. That is,

$$(\mathbf{M}^{(1)}, \Omega^{(1)}), ..., (\mathbf{M}^{(c)}, \Omega^{(c)}), ..., (\mathbf{M}^{(C)}, \Omega^{(C)}) \quad \overset{iid}{\sim} Pr(\mathbf{M}, \Omega|\mathbf{D}),$$

where $(\mathbf{M}^{(c)}, \Omega^{(c)})$ is the sample from the c-th iteration.

In this notation, $\mathbf{M}^{(c)} = \{M_{ij}^{(c)}\}_{1 \leq i \leq N, 1 \leq j \leq 2S-2, 1 \leq c \leq C}$ such that $M_{ij}^{(c)}$ is the substitution history of site i on branch j in the c-th iteration and has sufficient statistics that are $n_{abij}^{(c)}$ and $\phi_{aij}^{(c)}$. Note that $M_{ij}^{(c)}$ includes the number of substitutions from context a to nucleotide b at site i on branch j and the proportion of time site i on branch j has context a in the c-th iteration.

The 'substitution lengths' of changes from context a to nucleotide b on the phylogeny $\boldsymbol{\mu}_{ab}$ can be estimated by maximum likelihood estimates from $\mathbf{M}^{(c)}$. The maximum likelihood estimate

of $\boldsymbol{\mu}_{ab}$ for iteration c is

$$\widehat{\boldsymbol{\mu}}_{ab}^{(c)} = \begin{pmatrix} \widehat{\mu}_{ab1}^{(c)} \\ \vdots \\ \widehat{\mu}_{ab(2S-2)}^{(c)} \end{pmatrix}, \tag{4.10}$$

where

$$\widehat{\mu}_{abj}^{(c)} = \frac{\sum_i n_{abij}^{(c)}}{\sum_i \phi_{aij}^{(c)}} = \frac{N_{abj}^{(c)}}{\Phi_{aj}^{(c)}}. \tag{4.11}$$

And we can estimate $\boldsymbol{\mu}_{ab}$ by

$$\widehat{\boldsymbol{\mu}}_{ab} = \frac{1}{C}\sum_{c=1}^{C} \widehat{\boldsymbol{\mu}}_{ab}^{(c)}, \tag{4.12}$$

so that

$$\widehat{\mu}_{abj} = \frac{1}{C}\sum_{c=1}^{C} \widehat{\mu}_{abj}^{(c)}. \tag{4.13}$$

In Section 4.5, we will discuss alternative reasonable estimators of $\boldsymbol{\mu}_{ab}$.

By assuming the maximum likelihood estimates are asymptotically normally distributed, we approximate the variance of $\widehat{\mu}_{abj}^{(c)}$ for iteration c by the inverse Fisher information. This yields the variance estimate

$$-\frac{1}{\frac{d^2 \log p(\mathbf{n}_j, \boldsymbol{\phi}_j | \boldsymbol{\mu}_j, \mathbf{s}_{j0})}{d\mu_{ab}^2}}\Big|_{\widehat{\mu}_{abj}^{(c)}} = \frac{\widehat{\mu}_{abj}^{(c)2}}{N_{abj}^{(c)}} = \frac{N_{abj}^{(c)}}{\Phi_{aj}^{(c)2}} = \frac{\widehat{\mu}_{abj}^{(c)}}{\Phi_{aj}^{(c)}}. \tag{4.14}$$

By the Law of Total Variance, the variance of $\widehat{\mu}_{abj}$ is

$$Var(\widehat{\mu}_{abj}) = Var\left[E\left(\widehat{\mu}_{abj}|M^{(c)}\right)\right] + E\left[Var\left(\widehat{\mu}_{abj}|M^{(c)}\right)\right]. \tag{4.15}$$

The first term can be estimated from the sample variance of $\widehat{\mu}_{abj}^{(c)}$ and the second term can be estimated from the sample average of the inverse Fisher information, $\frac{\widehat{\mu}_{abj}^{(c)}}{\Phi_{aj}^{(c)}}$.

And we can estimate the variance-covariance matrix of $\widehat{\boldsymbol{\mu}}_{ab}$ by

$$\widehat{Cov}(\widehat{\boldsymbol{\mu}}_{ab}) = \begin{pmatrix} \widehat{Var}(\widehat{\mu}_{ab1}) & \widehat{Cov}(\widehat{\mu}_{ab1}, \widehat{\mu}_{ab2}) & \cdots & \widehat{Cov}(\widehat{\mu}_{ab1}, \widehat{\mu}_{ab(2S-2)}) \\ \widehat{Cov}(\widehat{\mu}_{ab2}, \widehat{\mu}_{ab1}) & \widehat{Var}(\widehat{\mu}_{ab2}) & \cdots & \widehat{Cov}(\widehat{\mu}_{ab2}, \widehat{\mu}_{ab(2S-2)}) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{Cov}(\widehat{\mu}_{ab(2S-2)}, \widehat{\mu}_{ab1}) & \widehat{Cov}(\widehat{\mu}_{ab(2S-2)}, \widehat{\mu}_{ab2}) & \cdots & \widehat{Var}(\widehat{\mu}_{ab(2S-2)}) \end{pmatrix}, \tag{4.16}$$

where

$$\widehat{Var}\left(\widehat{\mu}_{abj}\right) = \frac{1}{C-1}\sum_{c=1}^{C}\left(\widehat{\mu}_{abj}^{(c)} - \widehat{\mu}_{abj}\right)^2 + \frac{1}{C}\sum_{c=1}^{C}\frac{\widehat{\mu}_{abj}^{(c)}}{\Phi_{aj}^{(c)}}. \tag{4.17}$$

$$\widehat{Cov}\left(\widehat{\mu}_{abj}, \widehat{\mu}_{abj'}\right) = \frac{1}{C-1}\left[\sum_{c=1}^{C}\widehat{\mu}_{abj}^{(c)} \cdot \widehat{\mu}_{abj'}^{(c)} - \frac{1}{C}\sum_{c=1}^{C}\widehat{\mu}_{abj}^{(c)} \cdot \sum_{c=1}^{C}\widehat{\mu}_{abj'}^{(c)}\right]. \tag{4.18}$$

Note that some sources of uncertainty are unfortunately not included in the estimates. For instance, we do not include uncertainty due to using substitution histories that are not generated from a context-dependent model.

4.3.4 Divergence Time Estimation

Multidivtime is a Bayesian MCMC program for estimating divergence times on a known rooted phylogeny with a relaxed autocorrelated clock model[9, 10, 165]. Data sets consisting of multiple genes can be analyzed in Multidivtime by assuming a common set of divergence times but allowing independent rate trajectories for each gene. Multidivtime takes a two-step procedure to estimate species divergence times from multigene data sets. First, it estimates branch lengths via maximum likelihood and uses the curvature of the log-likelihood surface to estimate a variance-covariance matrix between the branch length estimates for each gene. Second, it adopts an MCMC procedure to sample divergence times and rates by approximating the likelihood surface with a multivariate normal distribution for each gene, which is determined by the branch length estimates and the variance-covariance matrix obtained in the first step.

In this study, we use Multidivtime for data sets where 'substitution lengths' vary among substitution types rather than data sets where branch lengths vary among genes. By sampling the substitution histories from PhyloBayes, the substitution lengths and the associated variance-covariance matrix can be estimated as described above for each kind of nucleotide substitution. Different kinds of nucleotide substitutions are treated by Multidivtime in the same way as it treats different genes that share the same set of divergence times. Just as Multidivtime allows the rate trajectories of different genes to independently vary over the phylogeny and just as it allows some genes to change rate in a more clock-like fashion than others, our analyses with Multidivtime have different substitution types change rate independently and allow some substitution types to be more clock-like than others. A weakness of Multidivtime that exists for multigene analyses is that it ignores the possibility of correlated rate changes among genes. Likewise, a weakness of our analyses of changing substitution rates over time is that correlated changes in rate among substitution types are biologically plausible but Multidivtime assumes the rates change independently. An additional shortcoming of using Multidivtime for studying substitution rate change is the treatment of variance and covariance structure of estimated

substitution lengths. Whereas Multidivtime can account for covariances in estimation error of substitution lengths among branches for each substitution type, its current implementation assumes no covariance in estimation error among among substitution types.

4.3.5 Data and Settings in PhyloBayes

We explored the clock-like fashion of CpG transitions using a subset of previously studied data set from Kim et al.[158]. We obtained sequence data from 9 species of primates orthologous to human Chromosome 16 (hg19.chr16:60842337-61130970) from Multiz alignments of 100 vertebrates in UCSC genome browser[123]. The nine species of primates analyzed are related through a phylogenetic tree in Figure 4.1 and bushbaby is the outgroup species. All the analyses are done with the fixed tree topology in Figure 4.1. These 9 species were selected because they yield a phylogeny for which branches are long enough to yield substantial information about substitution lengths but short enough to lessen the chance of multiple substitutions per site and avoid serious concerns about alignment uncertainty.

Because CpG islands are usually free of methylation and not hypermutable[23], the substitution process for these regions is not similar to CpG dinucleotides in the rest of the genome. Therefore, CpG islands with the following conditions are removed from the analysis: GC content of 50% or greater, length greater than 200 bp, observed/expected CpG content greater than 0.6. Similarly, positions in the exons and repetitive elements (identified by RepeatMasker[125]) are excluded from the analysis in the hope of obtaining neutrally evolving regions. The final data set contains approximately 115 Kb for each species and the proportion of CpG sites is around 1%.

We used the GTR [55] with four categories discrete gamma distributed rate model[66] to generate the substitution histories. The substitution process is characterized by a vector of parameters $\Omega = (\mathbf{l}, \mathbf{r}, \boldsymbol{\rho}, \boldsymbol{\pi})$. The branch lengths are noted $\mathbf{l} = \{l_j\}_{1 \leq j \leq 2S-2}$. Rate heterogeneity among sites is modeled by a discrete-gamma distribution[66], and $\mathbf{r} = \{r_i\}_{1 \leq i \leq N}$ are the relative rates for each site. $\boldsymbol{\rho} = \{\rho_{ab}\}_{1 \leq a,b \leq 4}$ is the relative exchangeability between states and define $\rho_{ab} = \rho_{ba}$ if $a > b$. The stationary probability of each state is $\boldsymbol{\pi} = \{\pi(a)\}_{1 \leq a \leq 4}$ such that $\sum_a \pi(a) = 1$.

The priors for parameters in Ω in PhyloBayes are specified as follows. Branch lengths $\mathbf{l}$ are i.i.d. from an exponential prior of mean λ, which itself has an exponential prior of mean 0.1. The discrete-gamma distribution of rate variation among sites $\mathbf{r}$ is parameterized by a shape parameter α, with an exponential prior of mean 1. The exchangeability parameters $\boldsymbol{\rho}$ are i.i.d. from an exponential of mean 1. And finally, the stationary probability of each nucleotide state $\boldsymbol{\pi}$ has a flat Dirichlet prior.

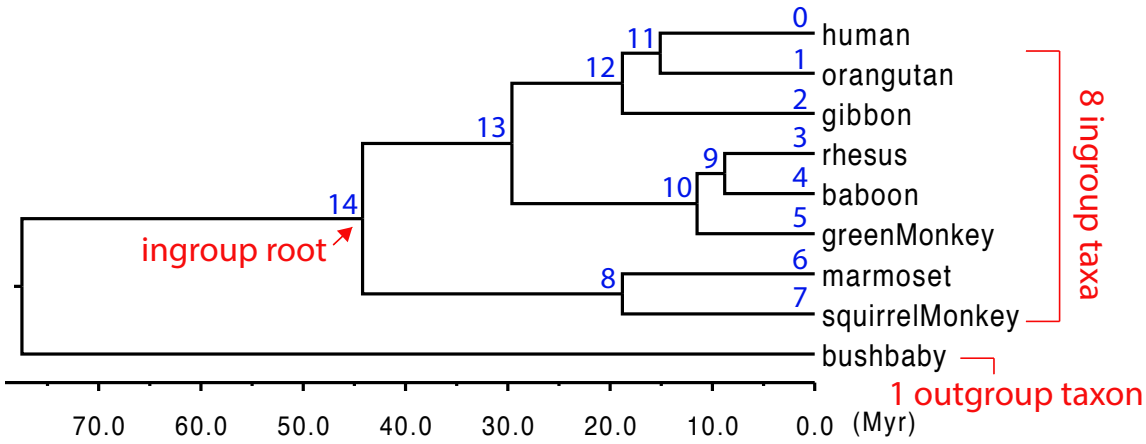


Figure 4.1: Phylogenetic tree for nine primate species used in the divergence time estimation analyses. There are 8 ingroup taxa with an outgroup species, bushbaby, to root the tree in the analyses. Nodes are labeled from 0 to 14. The "true" divergence times for nodes are obtained from TimeTree[4, 5] and indicated by the time line with time unit in one million years.

4.4 Results

4.4.1 Substitution Lengths for Each Type of Substitutions

We investigate the degree of deviation from clock-like behavior of each type of substitutions by first checking the substitution lengths of each type. If the rate of certain types of substitution is constant over time, then the root-to-tip substitution lengths for that type are expected to be similar for all lineages. The substitution lengths for context-dependent and context-independent substitution types are estimated from a sample of substitution histories from PhyloBayes. The root-to-tip substitution lengths for all lineages are computed and normalized so that the average root-to-tip substitution length within the same substitution type is 1. The normalized root-to-tip substitution lengths are summarized in Figure 4.2 and Figure 4.3. These figures were inspired by corresponding figures from Hwang and Green[157].

Among all lineages, the root-to-marmoset and root-to-squirrel monkey substitution lengths are always the greatest, regardless of substitution types. Also, the root-to-human and root-to-orangutan substitution lengths are always the smallest. Species within the same clade and therefore similar generation times and body sizes tend to have similar root-to-tip substitution lengths. This observation is consistent with the generation-time effect such that rates of lineages with relatively shorter generation times are elevated. Even though this phenomenon is relatively minor for CpG transitions (Type 9), it reveals that CpG sites are still subject to mutations

other than methylation-origin, e.g. replication errors.

The average root-to-tip substitution lengths for each substitution type before normalization imply the rate of each type of substitution. For context-dependent substitutions, Type 9 (CpG transitions) has the highest rate, while Type 3 (non-CpG T→A and A→T) and Type 4 (non-CpG T→G and A→C) have the lowest rate. Moreover, CpG sites are mutational hotspots so that the rates at CpG sites are accelerated. The CpG transition rate (Type9) is much higher than non-CpG transition rates (Type 5 and Type 6). Furthermore, CpG transversion rates (Type 7 and Type 8) are also much higher than non-CpG transversion rates (Type 1 and Type 2). Kong et al.[109] suggested that the high CpG transversion rate stems not only from hypermutable CpG sites, but also from mutational bias favoring mutations that decrease G+C content. As a matter of fact, CpG transversion rates are actually comparable to non-CpG transition rates. For context-independent substitutions, transition rates (Type 5 and Type 6) are higher than transversion rates (Type 1 - Type 4), while Type 5 (G→A and C→T) is the highest among all. This observation is consistent with the hypothesis that mutation is bias toward A+T content[166, 167].

The spread of the root-to-tip substitution lengths (quantified by the variance after normalization) reflects the degree of deviation from clock-like behavior. Among all context-dependent substitution types, Type 9 (CpG transitions) has the smallest variance after normalization, showing that it is the most clock-like. Following CpG transitions, Type 8 (CpG G→T and C→A) has the second smallest variance after normalization. This suggests that CpG sites are more clock-like than non-CpG sites, but the other transversion types for CpG site, Type 7 (CpG G→C and C→G), have a large normalized variance for the root-to-tip substitution lengths. In addition, Type 1 (non-CpG G→C and C→G) and Type 6 (non-CpG A→G and T→C) have relatively small variances, meaning they are more clock-like than other types. In fact, Type 1 (G→C and C→G) and Type 6 (A→G and T→C) are also the most clock-like among all context-independent single nucleotide substitutions. This result, nevertheless, does not agree with the observations of Kim et al.[158] that transversions exhibit less generation-time effect and were more clock-like than transitions.

4.4.2 Divergence Time and Substitution Rate Estimates

The divergence times and substitution rates of the phylogeny in Figure 4.1 were estimated by considering one substitution type per analysis and by jointly considering the collection of a strand-symmetric context-dependent subsetitution types. The divergence times were also estimated from Multidivtime and BEAST with conventional analyses that have all substitution types change rate in the same way on each branch. For the purposes of comparison, the root time prior for all analyses was tight and was gamma with mean 44.2 Myr and standard deviation 0.1

Myr. For all Multidivtime analyses, the lognormal autocorrelated relaxed clock model (ACLD) was employed. The root rate had a gamma prior with mean and standard deviation 0.0012. The actual marginal priors on divergence times were formed by the combination of the root time prior and a Dirichlet process. The autocorrelation parameter in the ACLD model assumed a gamma prior with mean and standard deviation 0.023, which is the inverse of the prior root time. For BEAST analyses, either the strict clock or the uncorrelated lognormal relaxed clock (UCLD) was selected. For the strict clock case, the single rate of the phylogeny had the same prior as the root rate prior in Multidivtime analyses, while for the UCLD case the mean rate had that prior. In addition to the root time prior, a birth-death process was assigned for the priors on divergence times in BEAST analyses. The mean growth rate had a uniform $(0, 100000)$ prior and the relative death rate had a uniform $(0, 1)$ prior. Rate heterogeneity among sites was modeled by a four-category discrete-gamma rate model with an exponential prior with mean 0.5 for the gamma shape parameter.

The divergence time estimates and 95% credible intervals are listed in Table 4.3 and Table 4.4. Because the root time had a very tight prior, the estimated root times are not reported here. For Bayesian divergence time estimation, all the information for separating rates and times comes from their prior distribution. Since there exists little knowledge about the patterns of rate change and hence the prior on rates poorly characterized the rate change pattern, substitutions that evolve at a relatively constant rate are expected to give better divergence time estimates. This prediction is indeed consistent with the divergence time estimates from using the CpG transitions only, which outperforms all other analyses in terms of precision and accuracy. On the other hand, substitution types that occurred less frequently (e.g. CpG transversions) would have greater uncertainty associated with the estimated substitution lengths, and thus their divergence time estimates have greater uncertainty as well.

The divergence time estimates from combined analysis for context-dependent substitutions are dominated by less clock-like non-CpG substitutions, because non-CpG sites occupy about 99% in the sequence alignment. If the proportion of clock-like substitution types in the sequence alignment increases, the divergence time estimates from the combined analysis would improve. The combined analysis assumed 9 genes (substitution types) and came with smaller uncertainty for divergence times than the Multidivtime analysis with the sequence data directly, even though they both presume the ACLD model. The BEAST analysis with a strict clock results in very tight 95% credible intervals for all times, but their posterior means depart from the "true" times in TimeTree[4, 5]. The posterior means of divergence times from BEAST UCLD analysis are similar with the strict clock analysis, but because it allows rates to vary among branches, the 95% credible intervals are wider.

For the purpose of demonstrating chronological substitution rate pattern for each type of substitution, we placed constraints on all divergence times on the phylogeny with the "true" time

and re-estimated the divergence times and substitution rates from Multidivtime by separately analyzing each type of context-dependent substitutions. The inferred chronological substitution rates for each branch and each substitution type are presented in Figure 4.4. The chronological substitution rates for CpG substitutions are higher than non-CpG substitutions. To better visualize the pattern of chronological substitution rates, the inferred rates are normalized so that within type average is 1 and shown in Figure 4.5. It shows that the inferred rates among branches are more constant for CpG transitions (Type 9), followed by Type 1, Type 6 and Type 8 substitutions. This is consistent with the variance of normalized root-to-tip substitution lengths.

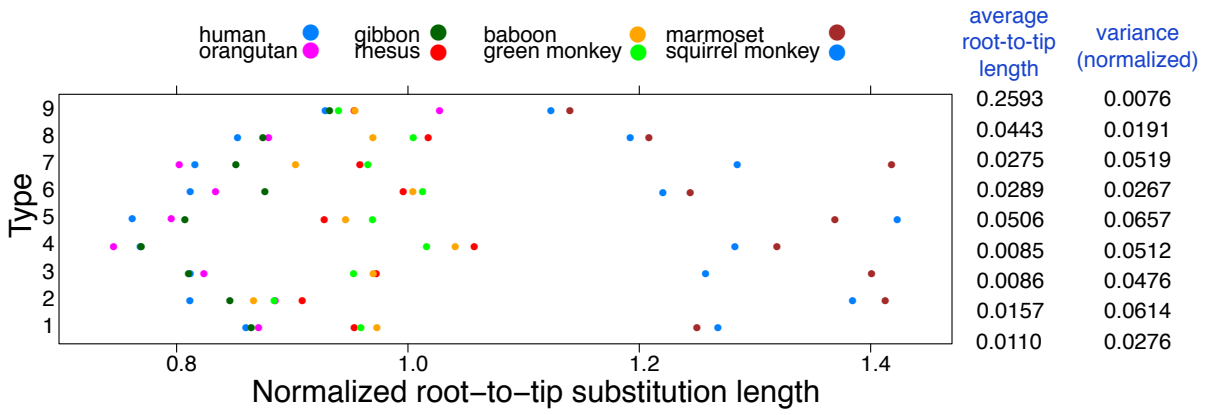


Figure 4.2: Normalized root-to-tip substitution lengths for each type of context-dependent substitutions. Types are defined in Table 4.2. Root-to-tip substitution lengths are normalized so that within type average substitution length is 1. The normalized root-to-tip substitution lengths for each lineage are labeled with different colors. The average root-to-tip substitution lengths for each type before normalization are reported as well as the variance after normalization.

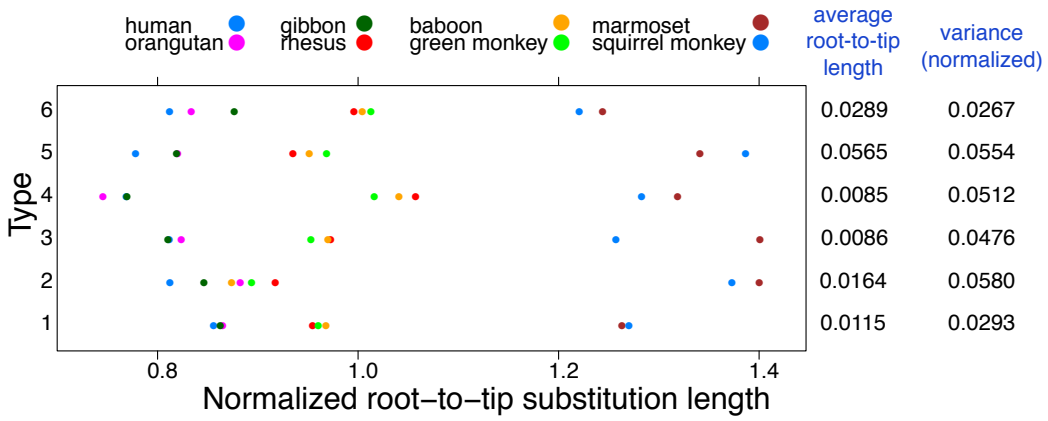


Figure 4.3: Normalized root-to-tip substitution lengths for each type of context-independent single nucleotide substitutions. Types are defined in Table 4.1. Root-to-tip substitution lengths are normalized so that within type average substitution length is 1. The normalized root-to-tip substitution lengths for each lineage are labeled with different colors. The average root-to-tip substitution lengths for each type before normalization are reported as well as the variance after normalization.

Table 4.3: Divergence time estimates for context-dependent substitutions. The means and the widths of the 95% posterior CIs of divergence times are reported. The nodes are labeled as in Figure 4.1 and the "true" divergence times are obtained from TimeTree[4, 5] and listed in Table 4.4. The types of the context-dependent substitutions are defined in Table 4.2. The third to the eleventh columns represent the results from Multidivtime and each type of context-dependent substitutions were analyzed separately. The last column represents the results from Multidivtime and all types of context-dependent substitutions were treated as different genes and analyzed jointly.

Node	"True" time	Type									Combined
		1	2	3	4	5	6	7	8	9	
8	18.8	17.63	16.92	17.61	18.73	17.24	18.97	15.27	15.69	17.48	17.42
		(14.99,20.69)	(14.02,21.03)	(14.98,20.95)	(15.75,22.66)	(14.73,20.56)	(16.91,21.44)	(8.07,25.46)	(9.62,23.46)	(14.44,20.71)	(16.5,18.36)
9	8.8	6.26	6.48	6.5	5.88	7.16	6.38	5.88	4.68	7.55	6.04
		(4.53,8.88)	(4.14,10.88)	(4.55,9.61)	(4.07,8.91)	(5.1,10.48)	(4.98,8.37)	(1.7,13.53)	(1.82,9.9)	(5.65,9.97)	(5.43,6.7)
10	11.5	9.37	9.58	9.99	10	10.72	10.28	8.09	5.85	11.7	9.43
		(7.06,12.94)	(6.36,15.46)	(7.3,14.2)	(7.25,14.32)	(7.93,15.01)	(8.31,13.06)	(3.11,17.04)	(2.56,11.98)	(9.13,14.84)	(8.6,10.33)
11	15.1	19.11	18.77	17.64	19.9	19.81	20.42	15.87	12.12	16.56	18.48
		(15.86,23.47)	(14.51,24.68)	(13.99,22.55)	(15.81,25.35)	(16.05,24.51)	(17.73,23.69)	(8.51,26.92)	(6.89,20.64)	(13.47,20.16)	(17.27,19.73)
12	18.8	21.3	20.14	19.85	22.22	21.62	22.7	17.31	13.53	18.98	20.46
		(17.89,25.69)	(15.7,26.13)	(15.97,24.89)	(17.94,27.68)	(17.74,26.35)	(19.89,26.07)	(9.66,28.6)	(7.94,22.38)	(15.67,22.78)	(19.17,21.75)
13	29.6	30.99	30.45	30.29	31.72	31.26	31.77	24.8	22.47	30.01	30.41
		(27.94,34.35)	(26.6,34.89)	(26.87,33.99)	(28.13,35.47)	(28.01,34.66)	(29.4,34.28)	(16,36.33)	(15.24,31.87)	(26.72,33.47)	(29.3,31.5)

Table 4.4: Divergence time estimates for single nucleotide substitutions. The means and the widths of the 95% posterior CIs of divergence times are reported. The nodes are labeled as in Figure 4.1 and the "true" divergence times are obtained from TimeTree[4, 5]. The types of the single nucleotide substitutions are defined in Table 4.1. The third to the eighth columns represent the results from Multidivtime and each type of single nucleotide substitutions were analyzed separately. The ninth column represents the results from Multidivtime and all types of context-dependent substitutions were treated as different genes and analyzed jointly. The last three columns represent the results from Multidivtime and BEAST (with a strict clock and with an uncorrelated lognormal relaxed clock) using the sequence alignment directly.

Node	"True" time	Type						Combined	Multidivtime	BEAST (clock)	BEAST (UCLD)
		1	2	3	4	5	6				
8	18.8	17.12	17.38	17.4	18.63	17.26	18.96	17.65	17.69	23.41	23.22
		(15.14,19.39)	(14.09,21.87)	(15.06,20.19)	(16.00,22.16)	(14.98,20.09)	(17.02,21.36)	(16.74,18.62)	(15.83,20.05)	(22.93,23.94)	(18.26,27.86)
9	8.8	6.11	7.02	6.34	5.73	7.27	6.38	6.17	6.41	5.44	5.58
		(4.82,8.05)	(4.36,12.00)	(4.68,8.84)	(4.15,8.37)	(5.4,10.03)	(5.12,8.42)	(5.56,6.90)	(5.00,8.51)	(5.22,5.67)	(4.3,7.02)
10	11.5	9.08	10.17	9.8	9.84	10.92	10.27	9.72	9.99	8.48	8.82
		(7.33,11.67)	(6.52,16.61)	(7.48,13.35)	(7.38,13.81)	(8.32,14.58)	(8.42,13.12)	(8.85,10.72)	(7.98,12.98)	(8.21,8.74)	(7.18,10.61)
11	15.1	19.1	19.58	17.54	19.98	19.51	20.48	19.15	19.34	13.13	13.88
		(16.57,22.46)	(14.86,25.66)	(14.3,21.54)	(16.23,24.63)	(16.09,23.62)	(17.91,23.84)	(17.92,20.45)	(16.51,22.81)	(12.69,13.48)	(10.91,16.83)
12	18.8	21.21	20.94	19.75	22.29	21.43	22.74	21.2	21.4	14.75	15.37
		(18.59,24.55)	(16.02,27.05)	(16.32,23.99)	(18.31,27.04)	(17.82,25.58)	(20.07,26.18)	(19.9,22.56)	(18.43,24.92)	(14.39,15.1)	(12.32,18.38)
13	29.6	30.68	30.81	30.19	31.7	31.18	31.78	30.93	31.02	24.38	24.36
		(28.28,33.37)	(26.69,35.19)	(27.19,33.39)	(28.48,35.05)	(28.09,34.2)	(29.55,34.28)	(29.86,31.98)	(28.5,33.69)	(23.9,24.81)	(20.96,27.97)

4.4.3 Simulation

To evaluate our method, especially the assumption that the distribution of substitution histories is robust to substitution model specification, we performed a simulation. A simulated data set was generated using the human sequence as the root sequence and the same tree topology from the real data set. The sequences were simulated using the estimated substitution lengths for each context-dependent substitution type on each branch. The simulated data set has about the same proportions of CpG, non-CpG C+G, and non-CpG A+T sites as the real data set. We then sampled the substitution histories from PhyloBayes and estimated the substitution lengths for each context-dependent substitution type as the real data analysis.

First, we used the estimated conventional branch lengths, which represent the expected number of substitutions per site on branches, to check if the substitution model is sensitive. We estimated the branch lengths from a sample of substitution histories by counting the number of substitutions on each branch. The branch lengths can also be estimated from other phylogenetic software (e.g. PAML[12]), and those estimates are similar to our estimates from sampling substitution histories (results not shown). The branch lengths of the real and the simulated data are similar and are listed in Table 4.5.

The root-to-tip substitution lengths for each context-dependent substitution type are compared between the real and the simulated data set in Figure 4.6. The substitution lengths are generally consistent between the real and the simulated data set, except that the root-to-tip substitution lengths for CpG transitions in the simulated data set are slightly less than those in the real data set. Since the CpG transition rate is significantly high, it may have multiple substitutions per site occurred when simulating the sequences, which are hard to be detected and led to smaller CpG substitution lengths. In general, it suggests that substitution histories are relatively insensitive to misspecification of the substitution model.

The divergence time estimates for the simulated data set are listed in Table 4.6. Similar to the real data analyses, the divergence time estimates for CpG transitions outperform other types of substitutions, with the exception in the divergence time between human and rhesus. This relatively deep divergence time on the phylogeny is vulnerable to multiple substitutions per site and it becomes difficult to estimate for CpG transition analysis.

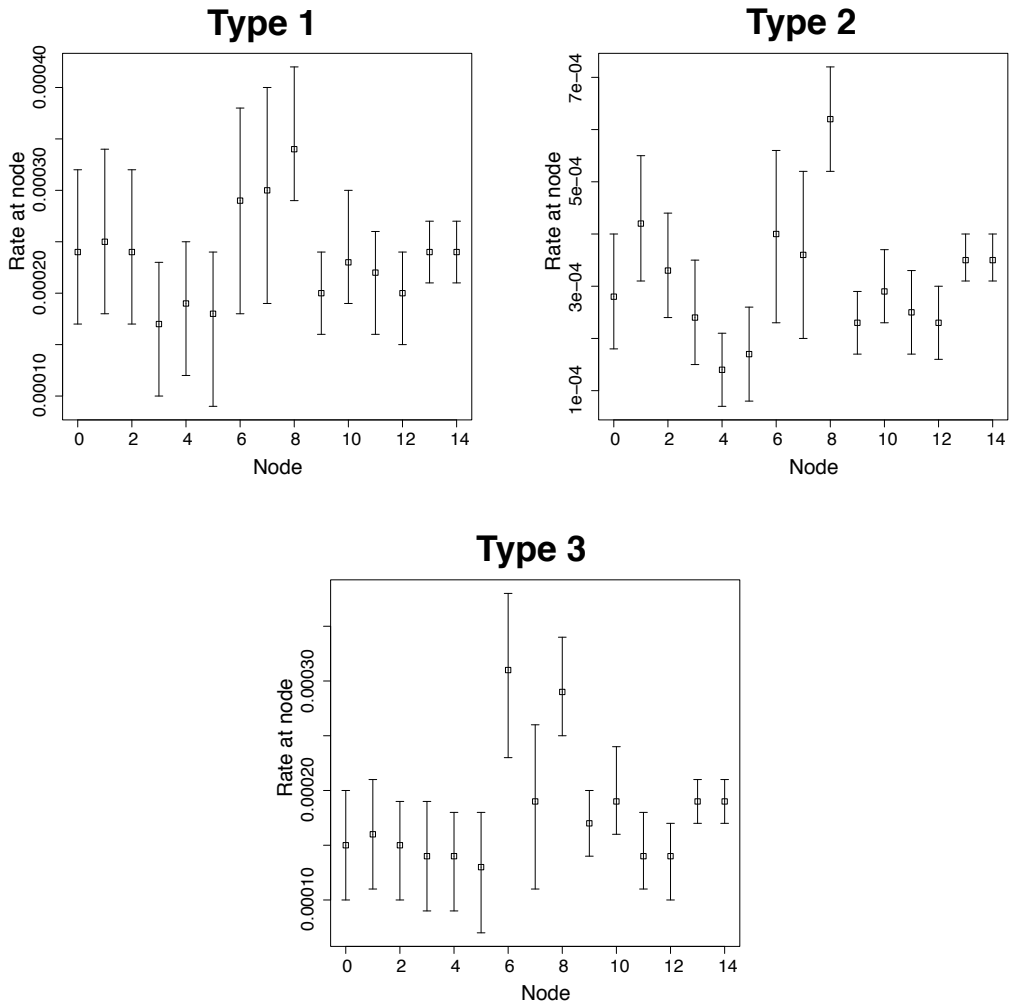


Figure 4.4: Estimated substitution rates at nodes for each type of substitutions with time constraints for all nodes. Rates were normalized so that within-type average is 1. Horizontal and vertical bars indicate 95% credible intervals.

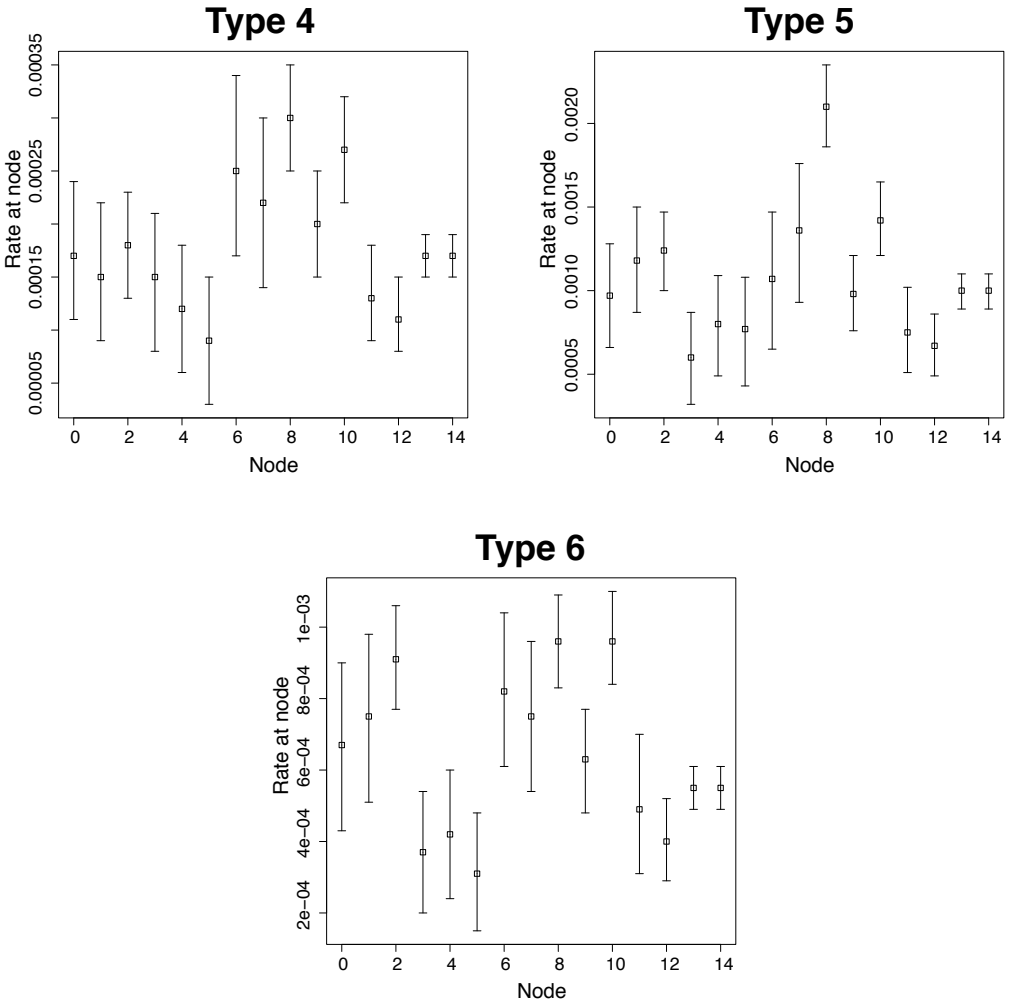


Figure 4.4: Estimated substitution rates at nodes for each type of substitutions with time constraints for all nodes. Rates were normalized so that within-type average is 1. Horizontal and vertical bars indicate 95% credible intervals.

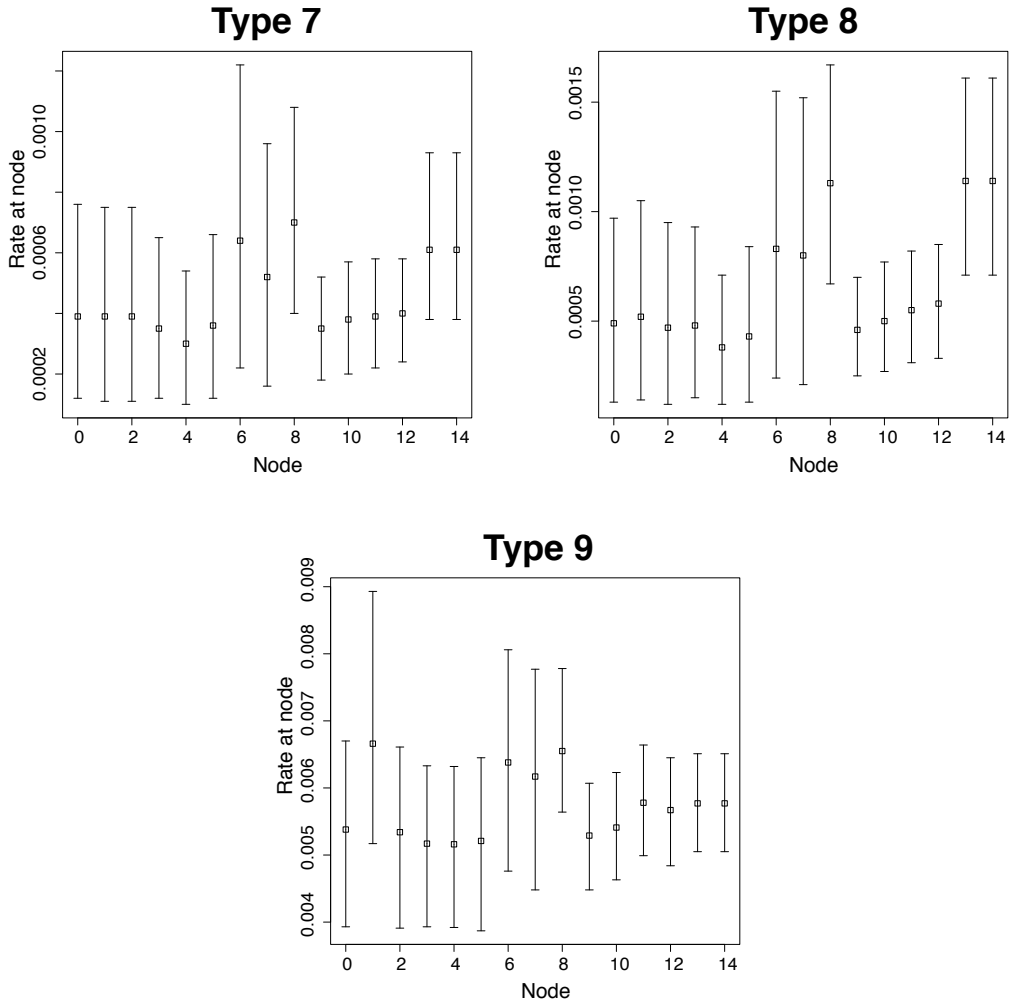


Figure 4.4: Estimated substitution rates at nodes for each type of substitutions with time constraints for all nodes. Rates were normalized so that within-type average is 1. Horizontal and vertical bars indicate 95% credible intervals.

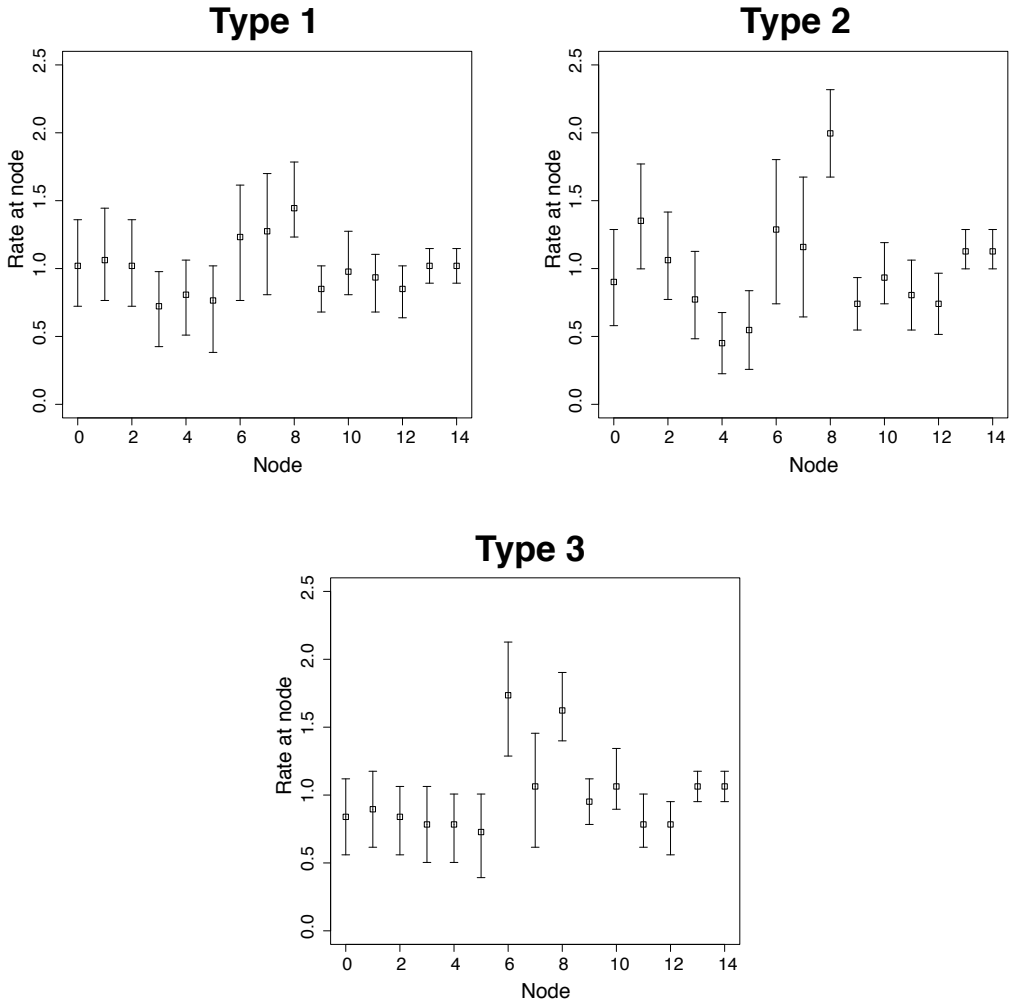


Figure 4.5: Estimated substitution rates at nodes for each type of substitutions with time constraints for all nodes. Rates were normalized so that within-type average is 1. Horizontal and vertical bars indicate 95% credible intervals.

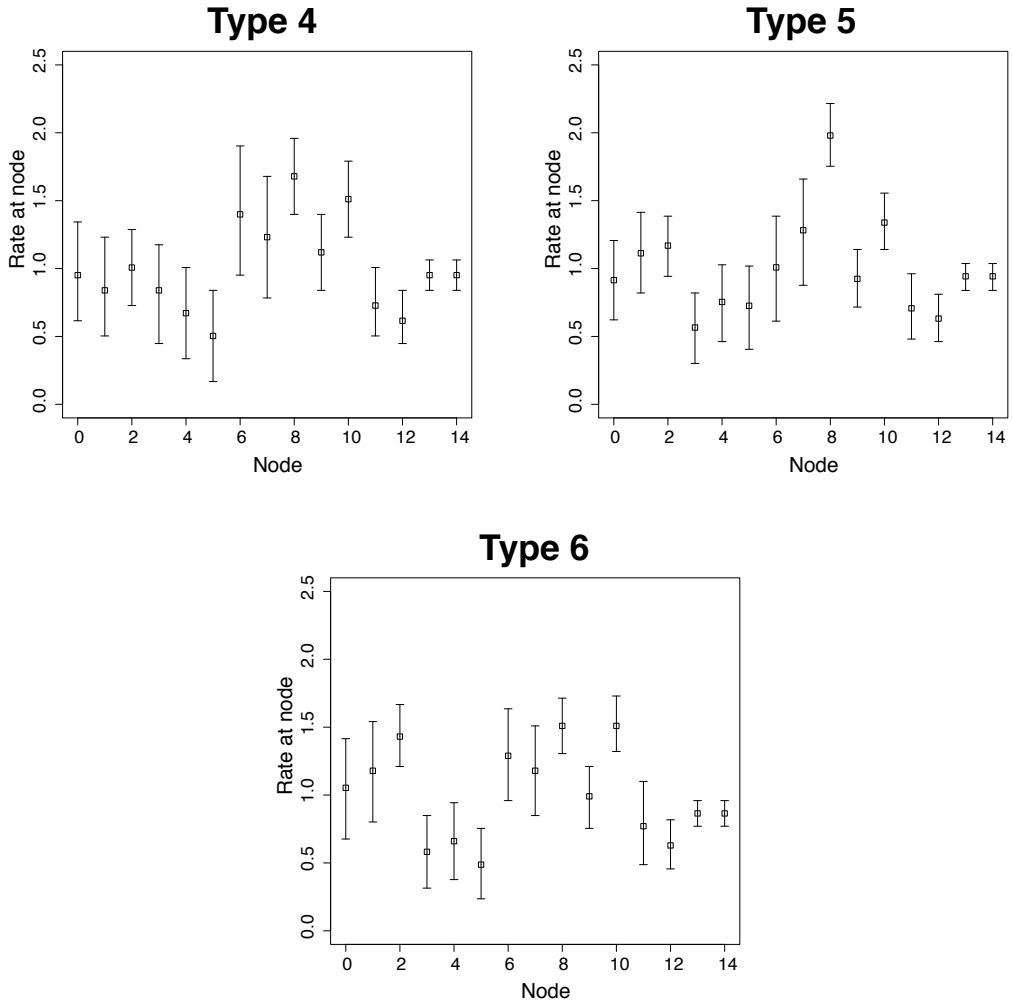


Figure 4.5: Estimated substitution rates at nodes for each type of substitutions with time constraints for all nodes. Rates were normalized so that within-type average is 1. Horizontal and vertical bars indicate 95% credible intervals.

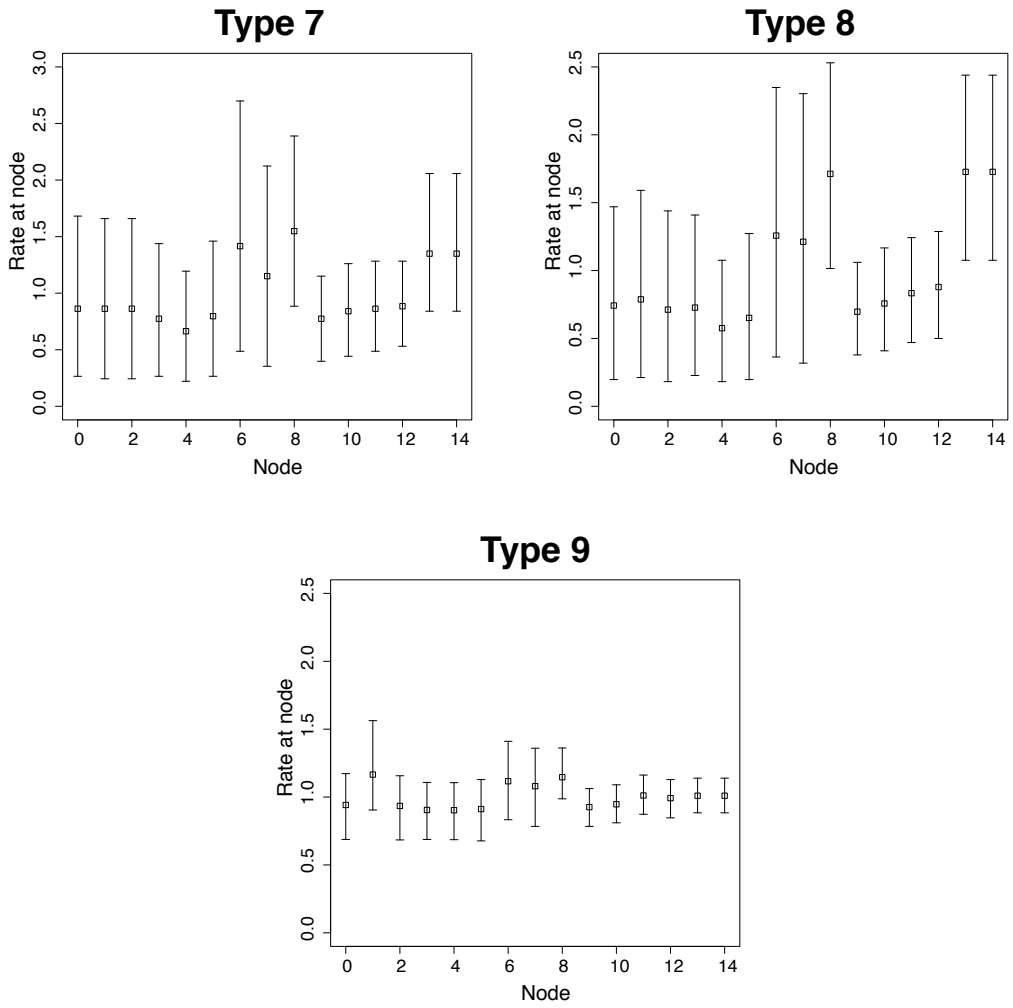


Figure 4.5: Estimated substitution rates at nodes for each type of substitutions with time constraints for all nodes. Rates were normalized so that within-type average is 1. Horizontal and vertical bars indicate 95% credible intervals.

Table 4.5: Comparison of the branch lengths from the real and the simulated data. The branches are labeled in Figure 4.1. The branch lengths are estimated from a sample of substitution histories by counting the number of substitutions for each branch.

Branch	0	1	2	3	4	5	6
Real data set	0.017334405	0.018901747	0.021634821	0.008157588	0.008328057	0.013420232	0.034580416
Simulated data set	0.016930604	0.018891276	0.021600704	0.007794052	0.008389608	0.013761355	0.036239078
Branch	7	8	9	10	11	12	13
Real data set	0.033825342	0.04274415	0.004759923	0.027628383	0.002071451	0.01109103	0.016625831
Simulated data set	0.033225726	0.044344769	0.004548446	0.028231608	0.002105073	0.010993123	0.016636274

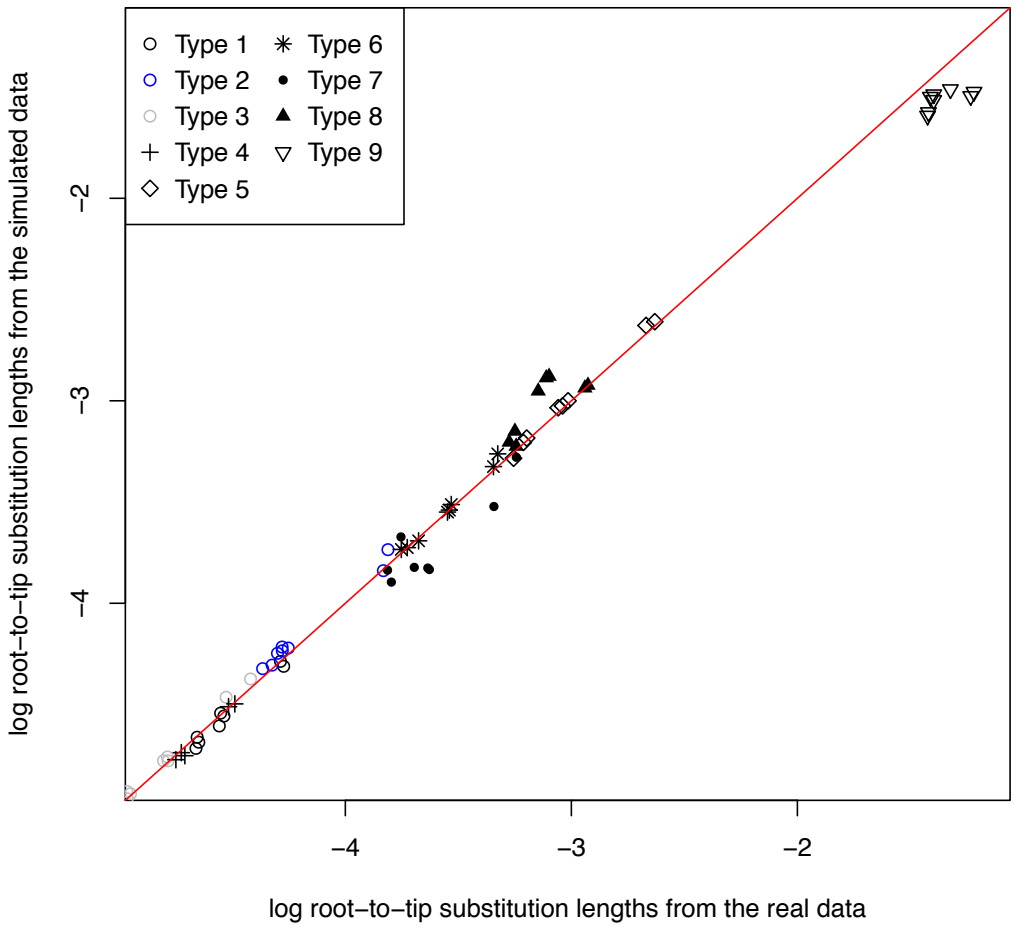


Figure 4.6: Comparison of the root-to-tip substitution lengths for each type of context-dependent substitutions between the real and the simulated data set. Substitution types are defined in Table 4.2 and labeled with different symbols. Each point represents the log root-to-tip substitution length in the real and the simulated data set for a particular substitution type. The substitution lengths are generally consistent between the real and the simulated data set, except that the root-to-tip substitution lengths for CpG transitions in the simulated data set are slightly less than those in the real data set.

Table 4.6: Divergence time estimates for context-dependent substitutions for the simulated data set. The means and the widths of the 95% posterior CIs of divergence times are reported. The nodes are labeled as in Figure 4.1. The types of the context-dependent substitutions are defined in Table 4.2. The third to the eleventh columns represent the results from Multidivtime and each type of context-dependent substitutions were analyzed separately. The last column represents the results from Multidivtime and all types of context-dependent substitutions were treated as different genes and analyzed jointly.

Node	Truth	Type									Combined
		1	2	3	4	5	6	7	8	9	
8	18.8	18.36	17.14	17.73	17.04	17.24	18.54	13.18	18.12	16.94	17.23
		(15.44,22.37)	(14.31,20.92)	(15.13,21.24)	(14.13,20.98)	(14.71,20.51)	(16.35,21.26)	(6.83,22.3)	(11.92,25.75)	(13.95,20.2)	(16.32,18.17)
9	8.8	7.33	6.37	6.33	5.52	6.99	6.79	6.5	4.14	7.61	6.18
		(5.07,11.38)	(4.07,10.61)	(4.37,9.63)	(3.74,8.57)	(5.02,10.13)	(5.18,13.91)	(1.64,15.58)	(1.49,10.06)	(5.61,10.21)	(5.52,6.9)
10	11.5	10.03	9.38	9.49	9.24	10.8	10.67	9.13	4.67	11.05	9.49
		(7.22,14.86)	(6.29,14.86)	(6.8,13.93)	(6.6,13.75)	(8.02,15.04)	(8.4,13.91)	(3.26,19.88)	(1.81,11.15)	(8.45,14.26)	(8.57,10.46)
11	15.1	19.7	19.33	17.24	20.82	20.46	20.17	16.8	9.36	15.66	18.76
		(15.89,24.99)	(15.14,24.75)	(13.47,22.57)	(16.67,26.2)	(16.64,24.98)	(17.13,23.82)	(8.76,27.75)	(3.94,19.24)	(12.63,19.31)	(17.46,20.09)
12	18.8	21.67	20.86	19.4	23.44	22.37	22.59	17.95	11.4	17.88	20.8
		(17.73,27.09)	(16.54,26.45)	(15.43,24.9)	(19.13,28.92)	(18.4,27.04)	(19.37,26.35)	(9.61,29.11)	(5.25,21.99)	(14.67,21.72)	(19.44,22.19)
13	29.6	31.78	31.93	29.24	31.99	31.95	31.84	28.4	21.73	25.83	30.64
		(28.47,35.66)	(28.31,35.72)	(25.7,33.31)	(28.38,35.79)	(28.71,35.23)	(29.19,34.57)	(18.79,39.02)	(14.36,31.13)	(22.41,29.54)	(29.29,31.78)

4.5 Discussion and Conclusions

Unfortunately, complete substitution histories and therefore the N_{abj} and Φ_{aj} values are not observed when phylogenetically related interspecific data sets are collected. Instead, only the sequences at the tips of a rooted evolutionary tree are observed. This lack of complete information makes the inference problem more challenging. Since allowing different substitution types to have different clocks is computationally daunting with context-dependent substitution, we provide another non-ideal solution. In this study, we collect substitution histories of homologous sequences according to their posterior distribution using context-independent substitution model via PhyloBayes[163]. These substitution histories are then used to infer context-dependent rates on each branch. Instead of different genes having different branch lengths, now we have different kinds of nucleotide substitutions with different 'substitution lengths'. The substitution types share the same set of divergence times, but each substitution types has its own rate trajectory on the phylogeny. The substitution lengths and the associated uncertainty for each substitution type can be approximated given the substitution histories. And finally, the divergence times and chronological substitution rates of each kind of substitutions are estimated with a relaxed molecular clock using another MCMC software, Multidivtime[9, 10, 165].

Another possible approach is to consider a Monte Carlo EM algorithm. Assuming the rooted tree topology is known, an expectation-maximization approach to maximizing the likelihood would proceed by specifying the initial guesses $\mu_{abj}^{(0)}$ for the parameter values. Thereafter, new parameter estimates $\mu_{abj}^{(k+1)}$ for iteration $k+1$ could be obtained by finding the values of μ_{abj} that maximize

$$E\left[\log p(\mathbf{n}_j, \boldsymbol{\phi}_j|\boldsymbol{\mu}_j, \mathbf{s}_{j0})|\mu_{abj}^{(k)}\right] = \sum_a \sum_b E\left[N_{abj}|\mu_{abj}^{(k)}\right] \log\{\mu_{ab}\} - E\left[\Phi_{aj}|\mu_{abj}^{(k)}\right] \mu_{ab}. \qquad (4.19)$$

However, the expectations $E[N_{abj}|\mu_{abj}^{(k)}]$ and $E[\Phi_{aj}|\mu_{abj}^{(k)}]$ are difficult to analytically determine. A Monte Carlo EM approach[168, 169] would instead sample a large number of substitution histories $\mathbf{M}^{(1)(k)}$, ..., $\mathbf{M}^{(C)(k)}$ from $Pr(\mathbf{M}|\mathbf{D}, \boldsymbol{\mu}_{ab}^{(k)})$. Each substitution history $\mathbf{M}^{(c)(k)}$ would be associated with its own values of the sufficient statistics $N_{abj}^{(c)}$ and $\Phi_{aj}^{(c)}$ for all branches. The sample means would then be substituted for the expectations in Equation 4.19 so that the values of $\mu_{abj}^{(k+1)}$ would maximize

$$\sum_a \sum_b \frac{\sum_{c=1}^{C} N_{abj}^{(c)}}{C} \log\{\mu_{ab}\} - \frac{\sum_{c=1}^{C} \Phi_{aj}^{(c)}}{C} \mu_{ab}.$$

Endpoint-conditioned sampling of substitution histories conditional on the observed data

has received increasing attention in statistical and molecular evolutionary literature[164, 170, 171, 162, 172, 173, 174, 175, 169, 160, 176, 177], but we instead choose to make a cruder approximation than would be obtained by Monte Carlo EM. The advantages of our approximation is feasibility of implementation and computation. Our approach is simply to sample mappings $\mathbf{M}^{(1)}$, ..., $\mathbf{M}^{(C)}$ from a simple and computationally tractable distribution and to then assume that these histories are actually sampled from $Pr(\mathbf{M}|\mathbf{D}, \boldsymbol{\mu}_{ab})$. We then estimate μ_{abj} as

$$\frac{1}{C}\sum_{c=1}^{C}\frac{N_{abj}^{(c)}}{\Phi_{aj}^{(c)}}.$$

While this estimator may be less promising than $\frac{\sum_c N_{abj}^{(c)}}{\sum_c \Phi_{aj}^{(c)}}$, the sampling distribution of our estimator can be asymptotically normal with its variance approximated by the law of total variance.

Our results indicate that CpG transitions occurred at a higher rate and are more clock-like than other types of substitutions in primates. The CpG transition analysis performs better than other analyses in estimating the primate divergence times. In addition, the degree to which other substitution types change rates over time can be diverse. It suggests that divergence time analyses should not assume that all substitution types change rates over time in the same way. Our approach allows other types of context-dependent rates not included in this study to be investigated. Future research direction includes sampling substitution histories directly from a context-dependent substitution model.

4.6 Appendix

The inferred substitution rates for each type of context-independent substitutions on each branch are presented in Figure 4.7. Complementary substitutions are placed side-by-side so that it is easier to check if combining complementary substitutions are reasonable. In general, the inferred substitution rates for complementary substitutions are alike except the inferred rate of G→C and C→G are different for some branches.

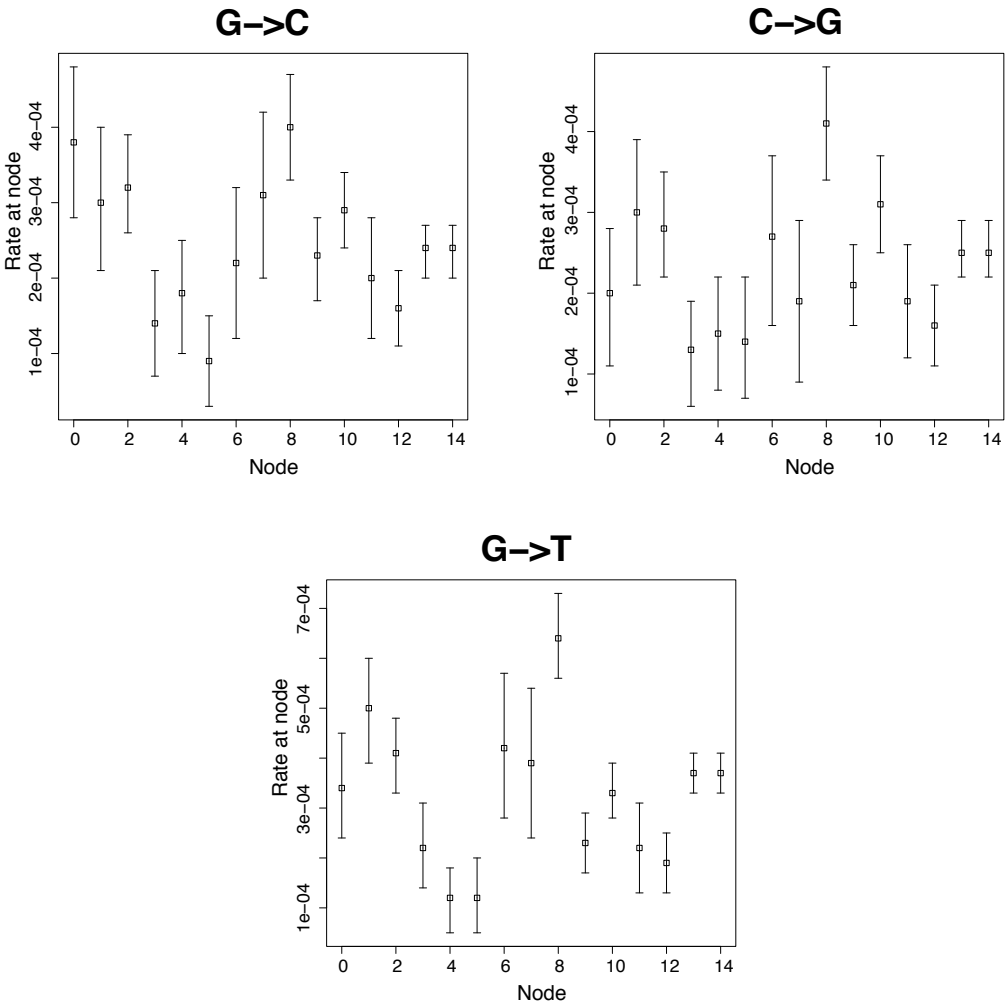


Figure 4.7: Estimated substitution rates at nodes for each type of single nucleotide substitutions with time constraints for all nodes. Horizontal and vertical bars indicate 95% credible intervals.

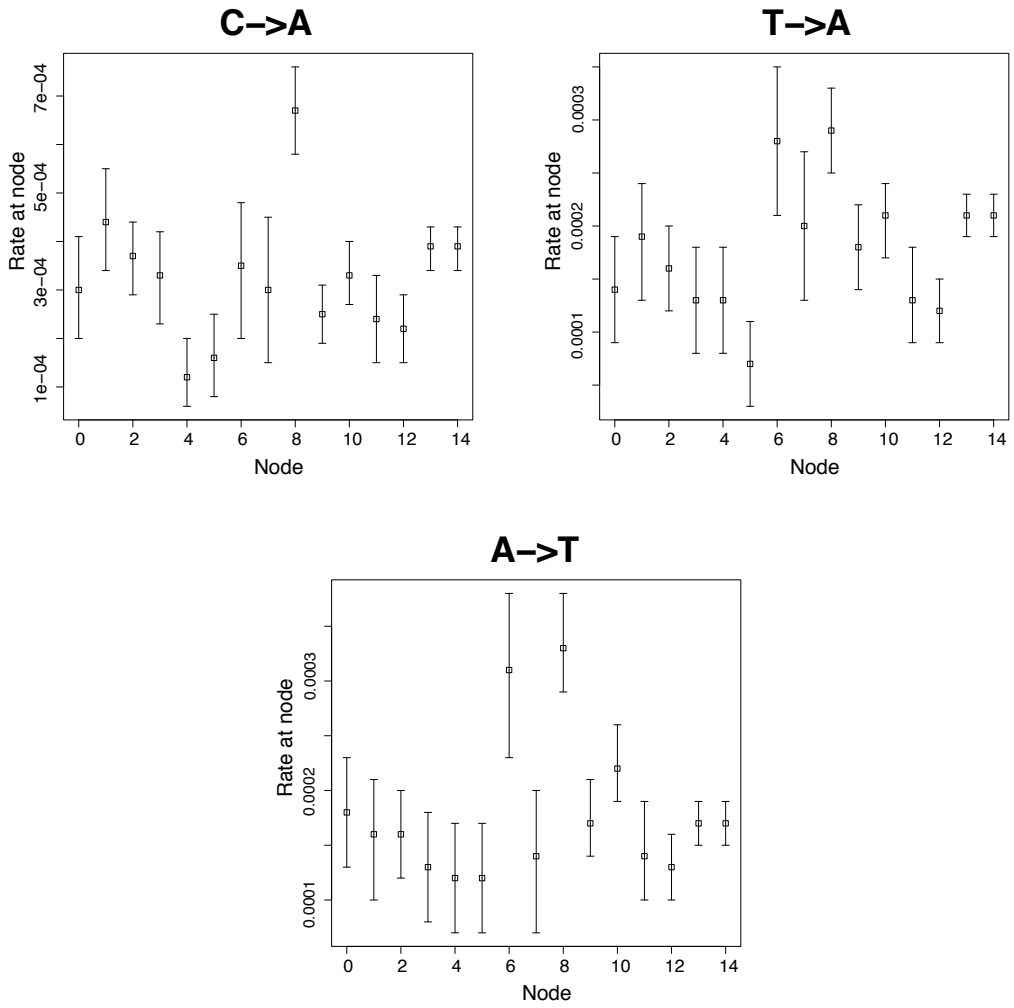


Figure 4.7: Estimated substitution rates at nodes for each type of single nucleotide substitutions with time constraints for all nodes. Horizontal and vertical bars indicate 95% credible intervals.

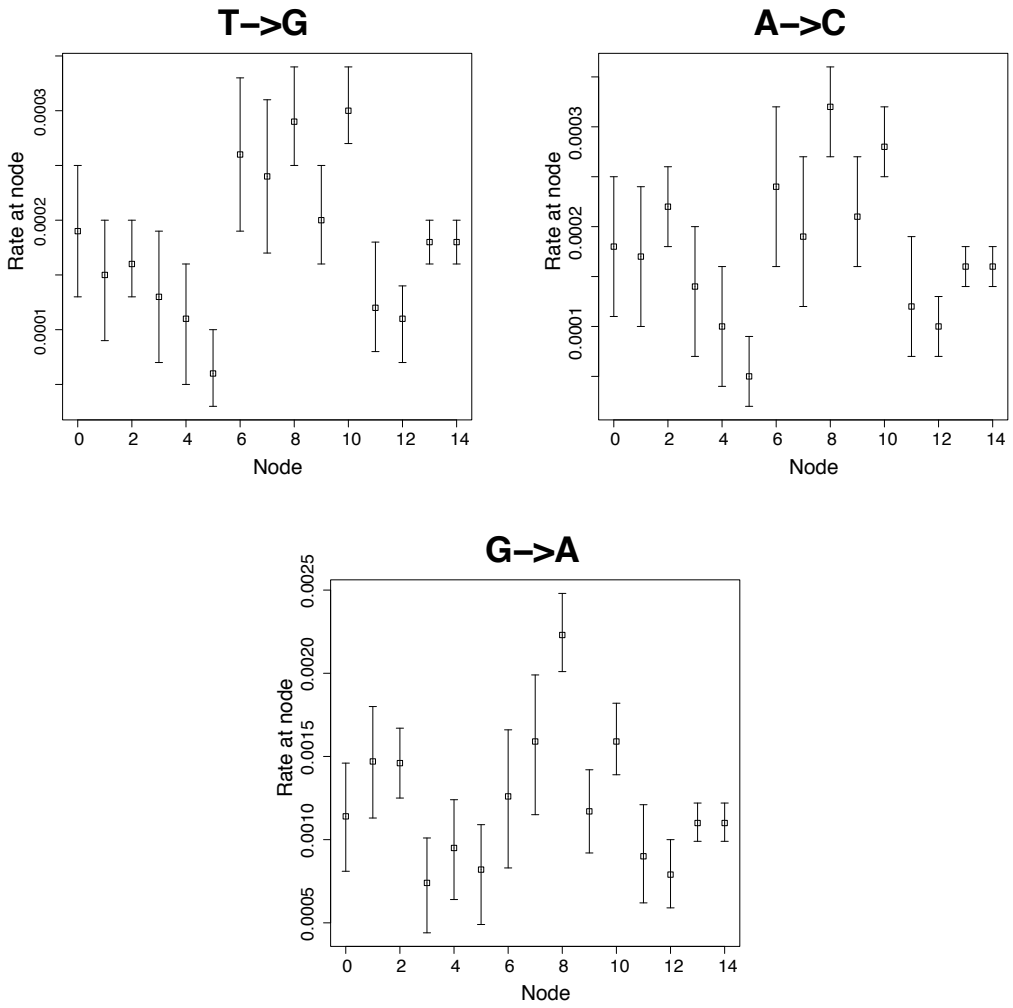


Figure 4.7: Estimated substitution rates at nodes for each type of single nucleotide substitutions with time constraints for all nodes. Horizontal and vertical bars indicate 95% credible intervals.

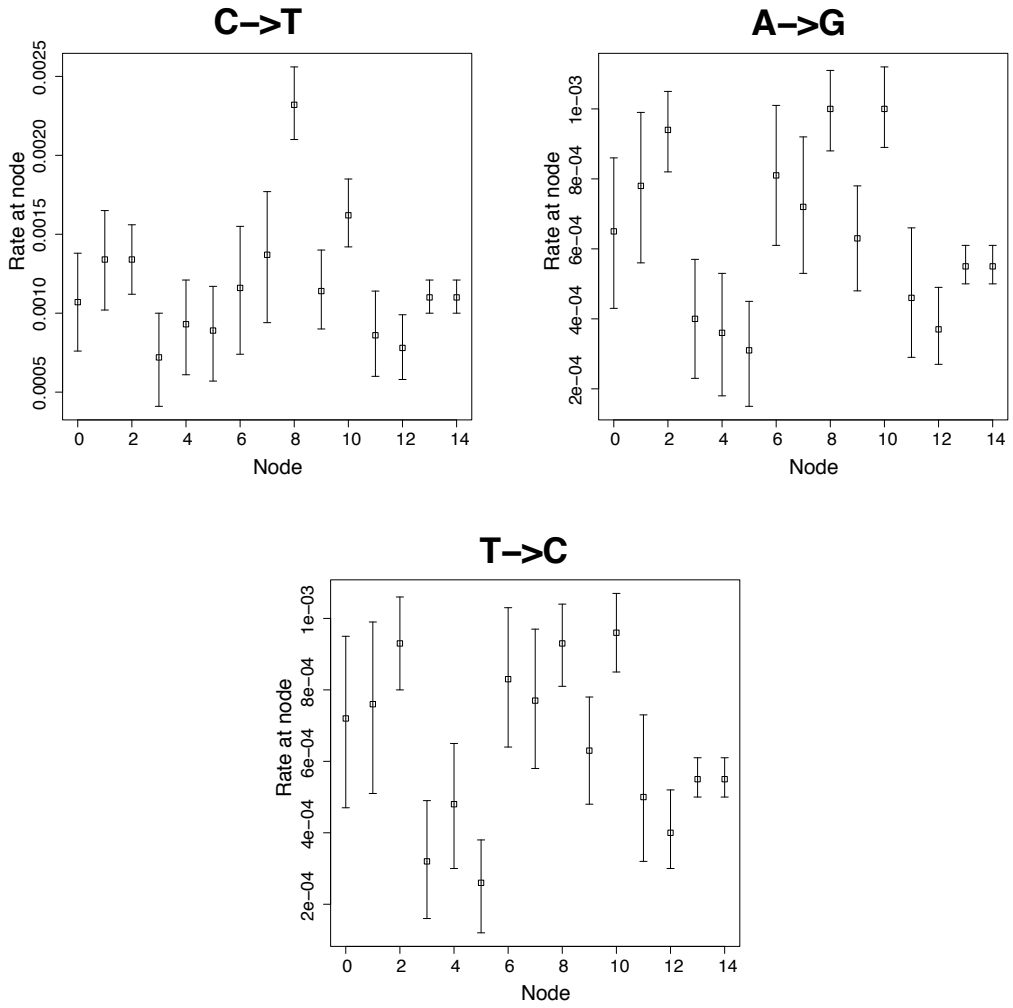


Figure 4.7: Estimated substitution rates at nodes for each type of single nucleotide substitutions with time constraints for all nodes. Horizontal and vertical bars indicate 95% credible intervals.

Chapter 5

Modeling Extensions

The mutation model implemented in Chapter 2 and Chapter 3 requires some assumptions about the mutation data or the sequence data. In this chapter, the mutation model is extended to relax some of the assumptions.

5.1 Incorporation of uncertainty in generation times in the mutation model

One of the possible extensions to the mutation model described in Chapter 2 is to relax the assumption of fixed known generation times.

Let R be branch rates and T be divergence times for the phylogenetic tree. Let X be the interspecific homologous aligned sequence and M be the mutation data on one branch. Moreover, we have one extra parameter g, which denotes the generation time of the branch with mutation data. The calculation of the likelihood of the aligned sequence data does not depend on generation time g.

$$P(R,T,g|X) \propto P(X|R,T)P(R,T,g) = P(X|R,T)P(R|T)P(T)P(g). \tag{5.1}$$

Then,

$$P(R,T|X) = \int_g P(R,T,g|X)dg \propto \int_g P(X|R,T)P(R|T)P(T)P(g)dg \tag{5.2}$$

$$= P(X|R,T)P(R|T)P(T)\int_g P(g)dg = P(X|R,T)P(R|T)P(T). \tag{5.3}$$

For cases where mutation data is present,

$$P(R,T,g|X,M) \propto P(X,M|R,T,g)P(R,T,g) = P(X|R,T)P(M|R,g)P(R|T)P(T)P(g). \tag{5.4}$$

Then,

$$P(R,T|X,M) = \int_g P(R,T,g|X,M)dg \propto \int_g P(X|R,T)P(M|R,g)P(R|T)P(T)P(g)dg \tag{5.5}$$

$$= P(X|R,T)P(R|T)P(T)\int_g P(M|R,g)P(g)dg. \tag{5.6}$$

Let us denote $R = \{R^*, R_C\}$, where R^* is the rate of the branch corresponding to tips with mutation data, and R_C are the branch rates of the rest. Assuming the number of mutations per site per generation y in a total of m bases screened follows a Poisson distribution, then

$$P(M|R,g) = P(M|R^*,g) = \frac{(m \times R^* \times g)^y \exp\{-m \times R^* \times g\}}{y!}. \tag{5.7}$$

If we assign a gamma prior for the generation time with a shape parameter α and a rate parameter β, then the prior density for the generation time is

$$P(g) = \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g}. \tag{5.8}$$

The prior mean and the variance of the generation time is $\frac{\alpha}{\beta}$ and $\frac{\alpha}{\beta^2}$, respectively. We can now simplify the integral in Equation (5.6) as follows.

$$\int_g P(M|R,g)P(g)dg = \int_g \frac{(m \times R^* \times g)^y \exp\{-m \times R^* \times g\}}{y!} \times \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} e^{-\beta g} dg \tag{5.9}$$

$$= \int_0^\infty \frac{(m \times R^*)^y \beta^\alpha}{\Gamma(y+1)\Gamma(\alpha)} \times g^{y+\alpha-1} e^{-(m \times R^* + \beta)g} dg \tag{5.10}$$

$$= \frac{(m \times R^*)^y \beta^\alpha}{\Gamma(y+1)\Gamma(\alpha)} \int_0^\infty g^{y+\alpha-1} e^{-(m \times R^* + \beta)g} dg \tag{5.11}$$

$$= \frac{(m \times R^*)^y \beta^\alpha}{\Gamma(y+1)\Gamma(\alpha)} \times \frac{\Gamma(y+\alpha)}{(m \times R^* + \beta)^{y+\alpha}} \tag{5.12}$$

$$= \frac{1}{y \times B(y,\alpha)} \times \frac{(m \times R^*)^y \beta^\alpha}{(m \times R^* + \beta)^{y+\alpha}} \tag{5.13}$$

In other words, we replace the Poisson model of the mutation data by a new model in Equation 5.13. It takes parameters: number of mutations per site per generation, branch rate R^*, total number of bases scanned, and α and β for specifying the prior distribution of the generation time.

5.2 Incorporation of false negative rate in the mutation model

The false negative rate refers to the mutations that have occurred but failed to be detected in the experiments to derive mutation rates. The uncertainty in the base-calling algorithm to count mutations can be incorporated in the mutation model described in Chapter 2.

Let p be the probability that a mutation that occurred is detected, and therefore the false negative rate is $1-p$. One way to model the false negative rate is to assume the number of mutations detected in the experiment is Poisson with mean equal to the product of the true mutation rate and the probability that a true mutation is detected. That is,

$$P(M|R,p) = \frac{(m \times R \times g \times p)^y \exp\{-m \times R \times g \times p\}}{y!}, \tag{5.14}$$

where R is the mutation rate per site per year, g is the generation time in years, and y is the number of mutations detected in a total of m bases assayed.

Another method to incorporate the false negative rate is as follows. The number of mutations Y that occurred in a total of m bases scanned is the sum of the mutations that occurred and were detected (true positives), y_{TP}, and the mutations that occurred but were not detected (false negatives), y_{FN}.

$$Y = y_{TP} + y_{FN}. \tag{5.15}$$

We assume the true positives y_{TP} is binomial so that the number of trials is Y and the probability of success in each trial is p. Then the likelihood of observing y mutations in a total of m bases scanned in the experiment is

$$Pr(y_{TP} = y|m,p) = \sum_{i=y}^{m} \binom{i}{y} (i-y)^{1-p} y^p \frac{\exp\{-m \times R \times g\}(m \times R \times g)^i}{i!} \tag{5.16}$$

$$= \sum_{i=y}^{m} \frac{1}{(i-y-1)!y!} \left(\frac{y}{i-y}\right)^p \exp\{-m \times R \times g\}(-m \times R \times g)^i. \tag{5.17}$$

This likelihood can replace the Poisson mutation likelihood but it is more computationally intensive.

5.3 Relaxing the neutral assumption of the mutation model by a mutation-selection balance model

The neutral assumption allows substitution rates and mutation rates to match so that mutation data could directly be informative for substitution rates. This assumption restricts our mutation model to be only applied to neutrally evolving regions. The mutation-selection balance model

(e.g. Halpern and Bruno[115]) provides a possible solution to relax the neutral assumption.

The rate of a particular nucleotide substitution R_{ij} from nucleotide i to j can be expressed as the product of the mutation rate, μ_{ij}, and the fixation probability, $Pr(Z_{ij})$. With low mutation rates, the fixation probability is a function of the product of the effective population size, N_e, and the relative fitness difference between nucleotides before and after the substitution, s[115].

$$Pr(Z_{ij}) \approx \frac{2s}{1 - \exp\{-2N_e s\}}.$$

And this product $N_e s$ can be inferred from interspecific sequence data. Therefore, we could link substitution rate to mutation rate by scaling the mutation rate with the fixation probability. Note that the mutation rate μ_{ij} is specified for nucleotides before and after the substitution in the mutation-selection balance model. Future mutation data containing information for each type of mutation might shed light on substitution rates of each kind through the mutation-selection balance model.

REFERENCES

[1] C. A. Russo, N. Takezaki, and M. Nei. Molecular phylogeny and divergence times of drosophilid species. *Molecular biology and evolution*, 12(3):391–404, May 1995.

[2] K. Tamura, S. Subramanian, and S. Kumar. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Molecular biology and evolution*, 21(1):36–44, January 2004.

[3] Darren J. Obbard, John Maclennan, Kang-Wook Kim, Andrew Rambaut, Patrick M. O'Grady, and Francis M. Jiggins. Estimating Divergence Dates and Substitution Rates in the Drosophila Phylogeny. *Molecular Biology and Evolution*, 29(11):3459–3473, November 2012.

[4] S. Blair Hedges, Joel Dudley, and Sudhir Kumar. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972, DEC 1 2006.

[5] Sudhir Kumar and S. Blair Hedges. TimeTree2: species divergence times on the iPhone. *Bioinformatics*, 27(14):2023–2024, JUL 15 2011.

[6] Soochin Cho, Suk-Won Jin, Adam Cohen, and Ronald E. Ellis. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. *Genome Research*, 14(7):1207–1220, July 2004.

[7] Drosophila 12 Genomes Consortium and AG. et al. Clark. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167):203–218, Nov 2007.

[8] Charles Darwin. *On the origin of species.* Cambridge University Press, Cambridge, UK;New York, 1859.

[9] J. L. Thorne, H. Kishino, and I. S. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Molecular biology and evolution*, 15(12):1647, 1998.

[10] H. Kishino, J. L. Thorne, and W. J. Bruno. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular biology and evolution*, 18(3):352–361, 2001.

[11] Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556, 1997.

[12] Z. Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

[13] A. J. Drummond and A. Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214–214, 2007.

[14] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*, 29(8):1969–1973, 2012.

[15] F. Ronquist, J. P. Huelsenbeck, M. Teslenko, P. Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, and M. A. Suchard. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542, 2012.

[16] E. Zuckerkandl and L. B. Pauling. *Molecular disease, evolution, and genetic heterogeneity.*, pages 189–225. Academic Press, New York, 1962.

[17] E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J . Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, 1965.

[18] MA Batzer, GE Kilroy, PE Richard, TH Shaikh, TD Desselle, CL Hoppens, and PL Deininger. Structure and variability of recently inserted alu family members. *nucleic Acids Research*, 18(23):6793–6798, DEC 11 1990.

[19] AS Yang, ML Gonzalgo, JM Zingg, RP Millar, JD Buckley, and PA Jones. The rate of CpG mutation in Alu repetitive elements within the p53 tumor suppressor gene in the primate germline. *Journal of molecular biology*, 258(2):240–250, MAY 3 1996.

[20] M Krawczak, EV Ball, and DN Cooper. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *American journal of human genetics*, 63(2):474–488, AUG 1998.

[21] M. W. Nachman and S. L. Crowella. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156:297–304, 2000.

[22] Sankar Subramanian and Sudhir Kumar. Neutral Substitutions Occur at a Faster Rate in Exons Than in Noncoding DNA in Primate Genomes. *Genome Research*, 13(5):838–844, May 2003.

[23] AP Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209–213, MAY 15 1986.

[24] J Josse, A Kornberg, and AD Kaiser. Enzymatic synthesis of deoxyribonucleic acid .8. frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *Journal of biological chemistry*, 236(3):864+, 1961.

[25] MN Swartz, A Kornberg, and TA Trautner. Enzymatic synthesis of deoxyribonucleic acid .11. further studies on nearest neighbor base sequences in deoxyribonucleic acids. *Journal of biological chemistry*, 237(6):1961–&, 1962.

[26] T Yomo and S Ohno. Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA CG deficiency TG CT excess. *Proceedings of the National Academy of Sciences of the United States of America*, 86(21):8452–8456, NOV 1989.

[27] K Jabbari and G Bernardi. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333:143–149, MAY 26 2004.

[28] M. Kimura and T Ohta. Probability of gene fixation in an expanding finite population. *Proceedings of the National Academy of Sciences of the United States Of America*, 71(9):3377–3379, 1974.

[29] M Kimura and T Ohta. Mutation and evolution at molecular level. *Genetics*, 73(APR):19–35, 1973.

[30] Motoo Kimura. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129):624–626, 1968.

[31] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[32] M Goodman. Role of immunochemical differences in phyletic development of human behavior. *Human biology*, 33(2):131–&, 1961.

[33] M Goodman, J Barnabas, G Matsuda, and GW Moore. Molecular evolution in descent of man. *Nature*, 233(5322):604–&, 1971.

[34] C. D. Laird, B. L. McConaughy, and B. J. McCarthy. Rate of fixation of nucleotide substitutions in evolution. *Nature*, 224(5215):149–154, 1969.

[35] DE Kohne, JA Chiscon, and BH Hoyer. Evolution of primate DNA sequences. *Journal of human evolution*, 1(6):627–644, 1972.

[36] RJ Britten. Rates of DNA-sequence evolution differ between taxonomic groups. *Science*, 231(4744):1393–1398, MAR 21 1986.

[37] A. P. Martin and S. R. Palumbi. Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, 90(9):4087–4091, 1993.

[38] C Schlotterer, B Amos, and D Tautz. Conservation of polymorphic simple sequence loci in Cetacean species. *Nature*, 354(6348):63–65, NOV 7 1991.

[39] T. Gojobori, WH Li, and D Graur. Patterns of nucleotide substitution in pseudogenes and functional genes. *Journal of molecular evolution*, 18(5):360–369, 1982.

[40] WH Li, CI Wu, and CC Luo. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *Journal of molecular evolution*, 21(1):58–71, 1984.

[41] MJ Sanderson. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular biology and evolution*, 14(12):1218–1231, DEC 1997.

[42] JP Huelsenbeck, B Larget, and D Swofford. A compound Poisson process for relaxing the molecular clock. *Genetics*, 154(4):1879–1892, 2000.

[43] AD Yoder and ZH Yang. Estimation of primate speciation dates using local molecular clocks. *Molecular biology and evolution*, 17(7):1081–1090, 2000.

[44] MJ Sanderson. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular biology and evolution*, 19(1):101–109, JAN 2002.

[45] S Aris-Brosou and ZH Yang. Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa. *Molecular biology and evolution*, 20(12):1947–1954, DEC 2003.

[46] A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):699–710, 2006.

[47] B. Rannala and Z. Yang. Inferring speciation times under an episodic molecular clock. *Systematic Biology*, 56(3):453–466, 2007.

[48] John Gillespie. *The Causes of Molecular Evolution*. Oxford University Press, New York, 1991.

[49] T H Jukes and C R Cantor. Evolution of protein molecules. In H N Munro, editor, *Mammalian protein metabolism*, pages 21–123. Academic Press, New York, 1969.

[50] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

[51] J. Felsenstein. *PHYLIP manual, version 3.2.*, 1989.

[52] J. Felsenstein and G. A. Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93–104, 1996.

[53] M. Hasegawa, T. Yano, and H. Kishino. A new molecular clock of mitochondrial dna and the evolution of hominiods. *Proc. Japan Acad.*, B(60):95–98, 1984.

[54] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[55] S. Tavaré. *Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences.*, volume 17, pages 57–86. Amer Mathematical Society, 1986.

[56] A. Rzhetsky and M. Nei. Tests of applicability of several substitution models for dna sequence data. *Molecular biology and evolution*, 12(1):131, 1995.

[57] Z. Yang, N. Goldman, and A. Friday. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular biology and evolution*, 11(2):316, 1994.

[58] S. Kumar. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics*, 143(1):537–548, 1996.

[59] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[60] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[61] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada. jmodeltest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8):772, 2012.

[62] Y. Fan, R. Wu, M. H. Chen, L. Kuo, and P. O. Lewis. Choosing among partition models in bayesian phylogenetics. *Molecular biology and evolution*, 28(1):523, 2011.

[63] W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M. H. Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160, 2011.

[64] G. Baele, P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and Alekseyenko A. V. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular biology and evolution*, 29(9):2157–2167, 2012.

[65] G. B. Golding. Estimates of dna and protein sequence divergence: an examination of some assumptions. *Molecular biology and evolution*, 1(1):125, 1983.

[66] Z. Yang. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, 1994.

[67] M. Hasegawa, H. Kishino, and T. Yano. Man's place in hominoidea as inferred from molecular clocks of dna. *Journal of Molecular Evolution*, 26(1-2):132–147, 1987.

[68] Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 193:993–1005, 1995.

[69] T. Uzzell and K. W. Corbin. Fitting discrete probability distributions to evolutionary events. *Science*, 172(3988):1089–1096, 1971.

[70] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3(5):418, 1986.

[71] L. Jin and M. Nei. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular biology and evolution*, 7(1):82, 1990.

[72] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512, 1993.

[73] Joseph Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic zoology*, 22(3):240–249, 1973.

[74] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[75] Z. Yang. Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Molecular biology and evolution*, 10(6):1396, 1993.

[76] D. V. Lindley. *Bayesian statistics : a review.* Philadelphia : Society for Industrial and Applied Mathematics, 1971.

[77] D. J. Poirier. Revising beliefs in nonidentified models. *Econometric Theory*, 14(4):483–509, 1998.

[78] J. P. Huelsenbeck, B. Larget, R. E. Miller, and F. Ronquist. Potential applications and pitfalls of bayesian inference of phylogeny. *Systematic Biology*, 51(5):673–688, 2002.

[79] Z. Yang and B. Rannala. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular biology and evolution*, 14(7):717, 1997.

[80] S. Li. *Phylogenetic tree construction using Markov chain Monte Carlo.* PhD thesis, Ohio State Univ., Columbus., 1996.

[81] S. Li, D. K. Pearl, and H. Doss. Phylogenetic tree construction using markov chain monte carlo. *Journal of the American Statistical Association*, 95(450):493–508, 2000.

[82] B. Mau and M. A. Newton. Phylogenetic inference for binary data on dendrograms using markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 6:122–131, 1997.

[83] B. Mau, M. A. Newton, and B. Larget. Bayesian phylogenetic inference via markov chain monte carlo methods. *Biometrics*, 55(1):1–12, 1999.

[84] Z. Yang and B. Rannala. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular biology and evolution*, 23(1):212–226, 2006.

[85] B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, 43(3):304–311, 1996.

[86] L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3):550–570, 1967.

[87] K. M. Pickett and C. P. Randle. Strange bayes indeed: uniform topological priors imply non-uniform clade priors. *Molecular phylogenetics and evolution*, 34(1):203–211, 2005.

[88] A. W. F. Edwards. Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society.Series B (Methodological)*, 32(2):155–174, 1970.

[89] B. F. Koop, M. Goodman, P. Xu, K. Chan, and J. L. Slightom. Primate eta-globin dna sequences and man's place among the great apes. *Nature*, 319(6050):234, 1986.

[90] H. Kishino and M. Hasegawa. Converting distance to time: Application to human evolution. *Methods in Enzymology*, 183:550–570, 1989.

[91] M. Hasegawa, H. Kishino, and T. Yano. Estimation of branching dates among primates by molecular clocks of nuclear dna which slowed down in hominoidea. *Journal of human evolution*, 18(5):461–476, 1989.

[92] A. J. Drummond and M. A. Suchard. Bayesian random local clocks, or one rate to rule them all. *BMC biology*, 8(1):114–114, 2010.

[93] T. Lepage, D. Bryant, H. Philippe, and N. Lartillot. A general comparison of relaxed molecular clock models. *Molecular biology and evolution*, 24(12):2669–2680, 2007.

[94] R. Bouckaert, J. Heled, D. Kühnert, T. G. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. BEAST2: A software platform for Bayesian evolutionary analysis, 2013.

[95] A. Rambaut and A. J. Drummond. Tracer v1.4, 2007.

[96] A Drummond and AG Rodrigo. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Molecular biology and evolution*, 17(12):1807–1815, DEC 2000.

[97] B Korber, M Muldoon, J Theiler, F Gao, R Gupta, A Lapedes, BH Hahn, S Wolinsky, and T Bhattacharya. Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288(5472):1789–1796, JUN 9 2000.

[98] A Rambaut. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4):395–399, APR 2000.

[99] T Leitner and J Albert. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proceedings of the National Academy of Sciences of the United States of America*, 96(19):10752–10757, SEP 14 1999.

[100] Elaine R. Mardis. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9:387–402, 2008.

[101] Jared C. Roach, Gustavo Glusman, Arian F. A. Smit, Chad D. Huff, Robert Hubley, Paul T. Shannon, Lee Rowen, Krishna P. Pant, Nathan Goodman, Michael Bamshad, Jay Shendure, Radoje Drmanac, Lynn B. Jorde, Leroy Hood, and David J. Galas. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*, 328(5978):636–639, APR 30 2010.

[102] Donald F. Conrad, Jonathan E. M. Keebler, Mark A. DePristo, Sarah J. Lindsay, Yujun Zhang, Ferran Casals, Youssef Idaghdour, Chris L. Hartl, Carlos Torroja, Kiran V. Garimella, Martine Zilversmit, Reed Cartwright, Guy A. Rouleau, Mark Daly, Eric A. Stone, Matthew E. Hurles, Philip Awadalla, and 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. *Nature genetics*, 43(7):712–U137, JUL 2011.

[103] C. Haag-Liautard, M. Dorris, X. Maside, S. Macaskill, D. Halligan, D. Houle, B Charlesworth, and Keightley. Direct estimation of per nucleotide and genomic deleterious mutation rates in drosophila. *Nature*, 445:82–85, 2007.

[104] Dee R. Denver, Peter C. Dolan, Larry J. Wilhelm, Way Sung, J. Ignacio Lucas-Lledo, Dana K. Howe, Samantha C. Lewis, Kazu Okamoto, W. Kelley Thomas, Michael Lynch, and Charles F. Baer. A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. *Proceedings of the National Academy Of Sciences of the United States Of America*, 106(38):16310–16314, SEP 22 2009.

[105] DR Denver, K Morris, M Lynch, and WK Thomas. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. *Nature*, 430(7000):679–682, AUG 5 2004.

[106] S. Benzer. On the topography of the genetic fine structure. *Proceedings of the National Academy of Sciences of the United States of America*, 47:403–415, 1961.

[107] K. H. Wolfe, P. M. Sharp, and W. H. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337:283–5, 1989.

[108] Giorgio Matassi, Paul M. Sharp, and Christian Gautier. Chromosomal location effects on gene sequence evolution in mammals. *Current Biology*, 9(15):786 – 791, 1999.

[109] Augustine Kong, Michael L. Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A. Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Wendy S. W. Wong, Gunnar Sigurdsson, G. Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Agnar Helgason, Olafur T. Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–475, August 2012.

[110] Simon Y. W. Ho, Robert Lanfear, Lindell Bromham, Matthew J. Phillips, Julien Soubrier, Allen G. Rodrigo, and Alan Cooper. Time-dependent rates of molecular evolution. *Molecular ecology*, 20(15):3087–3101, AUG 2011.

[111] JH Gillespie. The role of population size in molecular evolution. *Theoretical population biology*, 55(2):145–156, APR 1999.

[112] Asher D. Cutter. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. *Molecular biology and evolution*, 25(4):778–786, April 2008.

[113] N Metropolis, AW Rosenbluth, MN Rosenbluth, AH Teller, and E Teller. Equation of state calculations by fast computing machines. *Journal of chemical physics*, 21(6):1087–1092, 1953.

[114] WK Hastings. Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–&, 1970.

[115] AL Halpern and WJ Bruno. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7):910–917, JUL 1998.

[116] Jiaye Yu and Jeffrey L. Thorne. Dependence among sites in RNA evolution. *Molecular biology and evolution*, 23(8):1525–1537, AUG 2006.

[117] Jeffrey L. Thorne, Sang Chul Choi, Jiaye Yu, Paul G. Higgs, and Hirohisa Kishino. Population genetics without intraspecific data. *Molecular biology and evolution*, 24(8):1667–1677, AUG 2007.

[118] Sang Chul Choi, Asger Hobolth, Douglas M. Robinson, Hirohisa Kishino, and Jeffrey L. Thorne. Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular biology and evolution*, 24(8):1769–1782, AUG 2007.

[119] Ziheng Yang and Rasmus Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular biology and evolution*, 25(3):568–579, MAR 2008.

[120] Nicolas Rodrigue, Herve Philippe, and Nicolas Lartillot. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10):4629–4634, MAR 9 2010.

[121] F. Clemente and C. Vogl. Evidence for complex selection on four-fold degenerate sites in Drosophila melanogaster. *Journal of evolutionary biology*, 25(12):2582–2595, DEC 2012.

[122] David S. Lawrie, Philipp W. Messer, Ruth Hershberg, and Dmitri A. Petrov. Strong purifying selection at synonymous sites in ¡italic¿d. melanogaster¡/italic¿. *PLoS Genetics*, 9(5), 05 2013.

[123] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, Haussler, and D. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002.

[124] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian F A Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green,

David Haussler, and Webb Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4):708–15, April 2004.

[125] A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0, 1996-2004.

[126] Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.

[127] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: A web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, pages 19–10, 2010.

[128] Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb C Miller, W James Kent, and Anton Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455, 2005.

[129] Asher D. Cutter and Brian Charlesworth. Selection intensity on preferred codons correlates with overall codon usage bias in Caenorhabditis remanei. *Current biology*, 16(20):2053–2057, October 2006.

[130] Harris Todd W. et al. Yook, Karen. Wormbase 2012: more genomes, more data, new website. *Nucleic Acids Research*, 40:735–741, 2012.

[131] Federico Abascal, Rafael Zardoya, and Maximilian J. Telford. Translatorx: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research*, 38:7–13, 2010.

[132] Kazutaka Katoh and Daron M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, April 2013.

[133] P. Librado and J. Rozas. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11):1451–1452, June 2009.

[134] Nadia D. Singh, Amanda M. Larracuente, Timothy B. Sackton, and Andrew G. Clark. Comparative Genomics on the Drosophila Phylogenetic Tree. *Annual Review of Ecology, Evolution, and Systematics*, 40:459–480, 2009.

[135] Jack N. Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, 128(2):415–423, 2005.

[136] Shuichi Matsumura and Peter Forster. Generation time and effective population size in Polar Eskimos. *Proceedings of the royal society b-biological sciences*, 275(1642):1501–1508, JUL 7 2008.

[137] Serge A. Wich, S Suci Utami Atmoko, Tatang Mitra Setia, and Carel P. van Schaik. *Orangutans: Geographic Variation in Behavioral Ecology and Conservation*. OUP Oxford, 2008.

[138] Yun-Xin Fu. Estimating Mutation Rate and Generation Time from Longitudinal Samples of DNA Sequences. *Molecular Biology and Evolution*, 18(4):620–626, April 2001.

[139] Romke Koch, Henri G. A. M. van Luenen, Marieke van der Horst, Karen L. Thijssen, and Ronald H. A. Plasterk. Single nucleotide polymorphisms in wild isolates of Caenorhabditis elegans. *Genome Research*, 10(11):1690–1696, November 2000.

[140] Peter D. Keightley, Urmi Trivedi, Marian Thomson, Fiona Oliver, Sujai Kumar, and Mark L. Blaxter. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome research*, 19(7):1195–1201, JUL 2009.

[141] Kevin E. Langergraber, Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, Miho Inoue-Muruyama, John C. Mitani, Martin N. Muller, Martha M. Robbins, Grit Schubert, Tara S. Stoinski, Bence Viola, David Watts,

Roman M. Wittig, Richard W. Wrangham, Klaus Zuberbühler, Svante Pääbo, and Linda Vigilant. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39):15716–15721, September 2012.

[142] A. Scally and R. Durbin. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, 13:745–53, 2012.

[143] Julien Y. Dutheil, Ganesh Ganapathy, Asger Hobolth, Thomas Mailund, Marcy K. Uyenoyama, and Mikkel H. Schierup. Ancestral Population Genomics: The Coalescent Hidden Markov Model Approach. *Genetics*, 183(1):259–274, SEP 2009.

[144] Ann Gibbons. Turning back the clock: slowing the pace of prehistory. *Science*, 338(6104):189–191, 2012.

[145] Jaime E. Blair, Prachi Shah, and S. Blair Hedges. Evolutionary sequence analysis of complete eukaryote genomes. *BMC Bioinformatics*, 6, 2005.

[146] SYW Ho, MJ Phillips, AJ Drummond, and A Cooper. Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. *Molecular biology and evolution*, 22(5):1355–1363, MAY 2005.

[147] Stéphane Guindon. From trajectories to averages: an improved description of the heterogeneity of substitution rates along lineages. *Systematic Biology*, 62(1):22–34, 2013.

[148] Mario Dos Reis and Ziheng Yang. The unbearable uncertainty of bayesian divergence time estimation. *Journal of systematics and evolution*, 51(1):30–43, 2013.

[149] Alan Hodgkinson and Adam Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature reviews genetics*, 12(11):756–766, NOV 2011.

[150] Soojin V. Yi. Understanding neutral genomic molecular clocks. *Evolutionary biology*, 34(3-4):144–151, DEC 2007.

[151] JW Drake, B Charlesworth, D Charlesworth, and JF Crow. Rates of spontaneous mutation. *Genetics*, 148(4):1667–1686, APR 1998.

[152] AS Kondrashov and D Houle. Genotype-environment interactions and the estimation of the genomic mutation-rate in drosophila-melanogaster. *Proceedings of the royal society b-biological sciences*, 258(1353):221–227, DEC 22 1994.

[153] W Huang, BHJ Chang, X Gu, D HewettEmmett, and WH Li. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. *Journal of molecular evolution*, 44(4):463–465, APR 1997.

[154] LD Hurst and H Ellegren. Sex biases in the mutation rate. *TRENDS IN GENETICS*, 14(11):446–452, NOV 1998.

[155] Nicolas Lartillot and Frederic Delsuc. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, 66(6):1773–1787, JUN 2012.

[156] T Ohta. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96–98, 1973.

[157] DG Hwang and P Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States Of America*, 101(39):13994–14001, SEP 28 2004.

[158] Seong-Ho Kim, Navin Elango, Charles Warden, Eric Vigoda, and Soojin V. Yi. Heterogeneous genomic molecular clocks in primates. *PLOS GENETICS*, 2(10):1527–1534, OCT 2006.

[159] M. Peifer, J. E. Karro, and H. H. von Gruenberg. Is there an acceleration of the CpG transition rate during the mammalian radiation? *Bioinformatics*, 24(19):2157–2164, OCT 1 2008.

[160] Asger Hobolth and Eric A. Stone. Simulation from endpoint-conditioned, continuous-time markov chains on a finite state space, with applications to molecular evolution. *Annals of applied statistics*, 3(3):1204–1231, SEP 2009.

[161] Asger Hobolth and Jeffrey L. Thorne. Sampling and summary statistics of endpoint-conditioned paths in dna sequence evolution. In Ming-Hui Chen, Lynn Kuo, and Paul O. Lewis, editors, *Bayesian Phylogenetics: Methods, Algorithms, and Applications*. Taylor & Francis, 2014.

[162] Nicolas Lartillot. Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of computational biology*, 13(10):1701–1722, DEC 2006.

[163] Nicolas Lartillot, Thomas Lepage, and Samuel Blanquart. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *BIOINFORMATICS*, 25(17):2286–2288, SEP 1 2009.

[164] R Nielsen. Mapping mutations on phylogenies. *Systematic biology*, 51(5):729–739, SEP-OCT 2002.

[165] JL Thorne and H Kishino. Divergence time and evolutionary rate estimation with multilocus data. *Systematic biology*, 51(5):689–702, SEP-OCT 2002.

[166] N Sueoka. Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2653–2657, APR 1988.

[167] N Sueoka. Directional mutation pressure, selective constraints, and genetic equilibria. *Journal of molecular evolution*, 34(2):95–114, FEB 1992.

[168] GCG Wei and MA Tanner. A Monte-Carlo implementation of the EM algorithm and the poor mans data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, SEP 1990.

[169] Asger Hobolth. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *Journal of computational and graphical statistics*, 17(1):138–162, MAR 2008.

[170] A Hobolth and JL Jensen. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical applications in genetics and molecular biology*, 4, 2005.

[171] N Rodrigue, N Lartillot, D Bryant, and H Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347(2):207–217, MAR 14 2005.

[172] Nicolas Rodrigue, Herve Philippe, and Nicolas Lartillot. Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular biology and evolution*, 23(9):1762–1775, SEP 2006.

[173] L Mateiu and B Rannala. Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. *Systematic biology*, 55(2):259–269, APR 2006.

[174] Vladimir N. Minin and Marc A. Suchard. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of mathematical biology*, 56(3):391–412, MAR 2008.

[175] Nicolas Rodrigue, Herve Philippe, and Nicolas Lartillot. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. *Bioinformatics*, 24(1):56–62, JAN 1 2008.

[176] Asger Hobolth and Jens Ledet Jensen. Summary statistics for endpoint-conditioned continuous-time markov chains. *Journal of applied probability*, 48(4):911–924, DEC 2011.

[177] Paula Tataru and Asger Hobolth. Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains. *BMC bioinformatics*, 12, DEC 5 2011.