Journal of the Korean Data & Information Science Society
2020, **31**(3), 663–673

http://dx.doi.org/10.7465/jkdi.2020.31.3.663
한국데이터정보과학회지

Evaluation of ultrametric phylogenetic tree estimation[†]

Yujin Chung[1]

[1]Department of Applied Statistics, Kyonggi University

Received 9 March 2020, revised 5 April 2020, accepted 6 April 2020

Abstract

A phylogenetic tree is a tree-like graphical representation of the evolutionary history of a group of organisms. An ultrametric tree is a phylogenetic tree scaled to time, which is estimated from DNA sequences of the organisms. `PAUP*` is a popular software used to estimate diverse types of phylogenetic trees including ultrametric trees. An ultrametric tree includes the lengths of edges which are the major parameters of interest, and `PAUP*` provides three ways to parameterize the edge lengths. In this work, analyses of simulated and real data were conducted to examine the performance of ultrametric tree estimations by `PAUP*`. The three parameterizations, called `relAge`, `Thorne`, and `Rambaut`, were found to be correctly estimated in most cases, although the `relAge` parameterization resulted in tree estimation with zero-length branches less often than others. The `Rambaut` parameterization had higher accuracy in the tree structure estimation in the simulation.

Keywords: Parameterization, `PAUP*`, phylogenetic tree, ranked topology, ultrametric tree.

1. Introduction

The evolutionary history of a group of organisms can be repsented through a tree-like graph called a phylogenetic tree. Several distance matrix, maximum likelihood, and Bayesian methods have been developed to estimate phylogenetic trees from DNA sequences (Saitou and Nei, 1987; Huelsenbeck and Ronquist, 2001; Swofford, 2002; Felsenstein, 2004; Altekar *et al.*, 2004). In particular, maximum likelihood approaches have been applied to diverse organisms (Garrison *et al.*, 2016; Hosner *et al.*, 2015; Siepel, 2009) because of their desirable properties such as consistency (Rogers, 2001; Yang, 1995; Truszkowski and Goldman, 2016; RoyChoudhury *et al.*, 2015).

Like graphs which have been broadly used in Bayesian networks (Park, 2019) and decision trees (Jeon and Cho, 2019), a phylogenetic tree consists of vertices (or nodes) and edges (or branches). A node in a phylogenetic tree represents either an observed or unobserved status of an organism which is either extant or extinct, and edges connect vertices. Unlike graphs,

[†] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2018R1C1B5044541).

[1] Assistant professor, Department of Applied Statistics, Kyonggi University, Kwanggyosan-ro 154-42, Suwon 16227, Korea. E-mail: yujinchung@kgu.ac.kr

the lengths of edges represent evolutionary times and a set of parameters of interest. Both the tree structure and branch lengths are unknown parameters, and the tree is estimated using sequence information sampled at a number of genetic sites. One special case of a phylogenetic tree is the ultrametric tree, in which the evolutionary times from the most recent common ancestor (MRCA), called *root*, to the extant nodes are the same. Ultrametric trees often represent the so-called *molecular clock*, which states the same mutation rate across all lineages of the tree (Lee and Ho, 2016). Moreover, ultrametric trees of DNA sequences sampled from different chromosome location can be further analyzed to perform a population or species-level analysis (Chung and Hey, 2017).

In terms of estimating ultrametric trees, fewer branch-length parameters need to be estimated because of the condition of equal evolutionary times from the root to the extant nodes. There are different parameterizations of branch lengths for ultrametric trees. `PAUP*` (Swofford, 2002) is a popular software which provides maximum likelihood estimations of phylogenetic trees including ultrametric trees. To estimate an ultrametric tree, `PAUP*` implements three different parameterizations: `relAge, Thorne`, and `Rambaut`. It is desired that parameterizations should not affect the performance of estimator. However, parameterizations of edge lengths can determine geometric and statistical properties of a metric space of ultrametric trees (Gavryushkin and Drummond, 2016). Moreover, in practice, very short edges can be estimated as zero or negative values for several reasons, such as numerical errors. Although it is important to avoid zero or negative branch lengths, particulary when the estimated trees are used in a subsequent analysis, the performance of software estimating ultrametric trees with different parameterizations has not been yet examined.

In this study, simultions were conducted to compare the performance of the three parameterizations in terms of the estimation accuracy of edge lengths and tree structure called *ranked topology*. The simultions considered various edge lengths of an ultrametric tree structure, which were inferred by `PAUP*` with the three parameterizations. In particular, the performance of estimations of very short edges was investigated.

The remainder of this paper is organized as follows. In section 2, the terminologies for ultrametric phylogenetic trees, and the standard evolutionary models and maximum likelihood approaches are introduced, as are the three edge-length parameterizations. Sections 3 and 4 describe the results of simulation studies and real data analyses, respectively. The conclusion is presented in Section 4.

2. Statistical evolutionary models

2.1. Terminologies

Consider a graph G defined as a pair (V, E) consisting of a finite set V of vertices and a set E of edges whose element is defined as $e = \{\{x, y\} | x, y \in V\}$ (Lauritzen, 1996). The degree of a vertex v of graph G is defined as the number of edges directly connected to the vertex v. A phylogenetic tree is defined as $\tau = (T, \mathbf{t})$, consisting of $T = (V, E)$, a graph, and $\mathbf{t}$, a vector of lengths of edges of T. If the following conditions are satisfied but not limited, a graph is called phylogenetic tree, or specifically, a *binary tree* or *bifurcating tree* (Semple and Steel, 2003).

(i) it has no cycles;

(ii) the degree of a vertex of T is smaller than or equal to 3. In particular, a vertex of degree one is called a *leaf*, *tip* or *terminal vertex*, and one of degree three is called an *internal vertex*;

(iii) T has at most one vertex of degree two. If T has a vertex of degree two, the vertex is called a *root*, and T is called a *rooted tree*. Otherwise, T is called an *unrooted tree*.

For phylogenetic trees, vertices and edges can also be called *nodes* and *branches*, respectively. Terminal nodes of phylogenetic tree T typically represent the observed states (e.g., nucleotides) of extant organisms, while internal nodes represent the unobserved states of their ancestors. Because an edge of T connects between a descendant and its ancestor, the length of an edge, if presented, represents the evolutionary times in terms of the number of generations, the number of mutations, or etc. The node directly connected to its descendant node is conventionally called a *"parent"* node, while that directly connected to its ancestral node is called a *"child"* node. A typical phylogenetic tree τ has terminal nodes with distinct labels and unlabelled internal nodes.

If the sums of edge lengths from the root to the tips are the same, the tree is called an ultrametric tree. For example, Figure 2.1 (a) shows an example of a ultrametric tree. The distances from the root to the tips of the tree are the same as t_3, which is called the *age* of the root node. The nodes with ages t_2 and t_1 are the MRCAs of organisms 1 and 2, and of organisms 3 and 4, respectively. Similarly, the root is the MRCA of organisms 1-4. A rooted tree $\tau = (T, \mathbf{t})$ with four tips has six edges, and hence the branch length vector $\mathbf{t}$ has six elements. For an ultrametric tree, while it is known that the tree has six edges, it is necessary to estimate the ages of the internal nodes rather than the lengths of all edges. Therefore, for an ultrametric tree, $\mathbf{t}$ in $(T, \mathbf{t})$ denotes a vector of the ages of the internal nodes. For example, $\mathbf{t} = (t_1, t_2, t_3)$.

A phylogenetic tree $\tau = (T, \mathbf{t})$ without branch length information is called the *topology*, T. The topologies of Figure 2.1 (b)-(c) are essentially identical, as the order of node ages is not relevant to topology. If a topology carries the relative order of ages of internal nodes, then the tree is called *ranked topology*. From the ranked tree in Figure 2.1 (b), the lengths of edges are unknown. However, it carries the information that the MRCA of tips 1 and 2 is older than that of tips 3 and 4. Therefore, the ranked tree of Figure 2.1 (a) is depicted in Figure 2.1 (b) but not Figure 2.1 (c).

2.2. Evolutionary models and maximum likelihood approach

We consider n aligned DNA sequences whose columns are called *sites*. The length of an alignment is defined as the number of sites. For example, Figure 3.1 shows four aligned sequences of length 10. Let D present n aligned DNA sequences of length L. Each site $\mathbf{d}_i = (d_{i1}, \ldots, d_{in})'$ contains a nucleotide ($d_{ij} \in \{A, C, G, T\}$) from each of the n sequences for $i = 1, \ldots, L$.

The standard likelihood-based approaches model the evolutionary process in terms of a phylogenetic tree $\tau = (T, \mathbf{t})$ and a substitution rate matrix Q, where T is a tree topology and $\mathbf{t}$ is a vector of branch lengths (Felsenstein, 2004; Yang, 2006). Because the present study focuses on ultrametric tree estimation, $\mathbf{t}$ is a vector of node ages. Substitutions at a particular site are described by a continuous-time Markov chain whose states represent the four nucleotides. A rate matrix Q defines the rates for a continuous-time Markovian

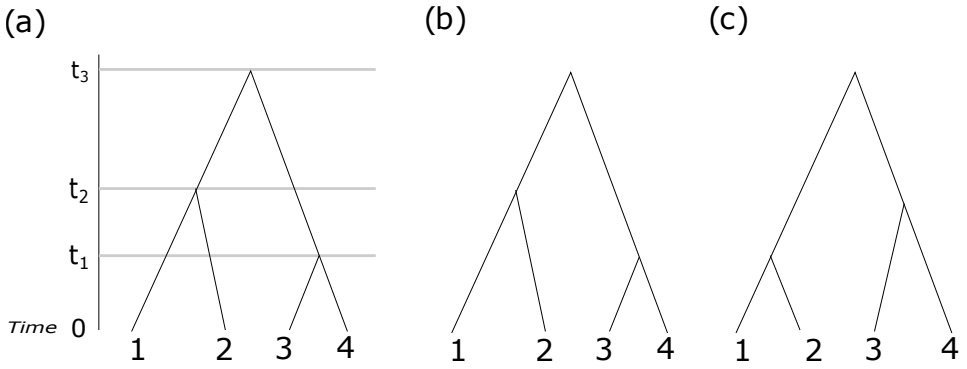

Figure 2.1 (a) Example of an ultrametric tree; (b) Ranked tree topology of the ultrametric tree in (a); (c) Example of another ranked tree topology; If the age orders of nodes in (b) and (c) are unknown, both represent an identical topology of the tree in (a).

nucleotide substitution process along each branch of τ. The Jukes-Cantor (JC) model (Jukes and Cantor, 1969) is one of the simplest substitution models:

$$Q = \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}, \tag{2.1}$$

where each nucleotide has the same rate λ of changing into another nucleotide. More general substitution models such as HKY (Hasegawa *et al.*, 1985), and GTR (Tavaré, 1986) have been developed, and they are time-reversible (Felsenstein, 2004). The resulting transition probability matrix is $\mathbf{P}(t) = e^{tQ} = \{p(i,j|t)\}$, where $p(i,j|t)$ is the probability of nucleotide i changing to j over time t for $i, j \in \{A, C, G, T\}$. Several methods for substitution model selection were developed (Nascimento *et al.*, 2017).

1 TACGTATTTT
2 TACGTATCAT
3 TACTTATCTT
4 TACTTATCTT

Figure 2.2 Example of a DNA alignment of four sequences of length 10

A rooted tree with n tips has $n-1$ internal nodes, $v_1, \ldots, v_{n-1}$, whose unobserved states are A, C, G, or T. We let $v_n, \ldots, v_{2n-1}$ denote the observed states of n terminal nodes. Let v_1 be the root node of the tree, then a site-likelihood is

$$L(\tau|\boldsymbol{d}_i) = \sum_{v_1,\ldots,v_{n-1}\in\{A,C,G,T\}} \cdots \sum \pi_{v_1} \left\{ \prod_{j=2}^{2n-1} p(anc(v_j), v_j|b_j) \right\},$$

where $anc(v_j)$ is the parent node of v_j, b_j is the length of the edge connecting v_j and $anc(v_j)$, and $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ is the stationary distribution of the nucleotides. Felsenstein (1981) provides an algorithm for computing the-site likelihood.

Because the sites of an alignment are often assumed to evolve independently of each other, the full likelihood function is as follows:

$$L(\tau|D) = \prod_{i=1}^{L} L(\tau|\boldsymbol{d}_i). \tag{2.2}$$

The maximum likelihood estimation of an ultrametric tree and the substitution rate parameters are obtained by maximizing the full likelihood function (2.2) subject to $0 < t_1 < t_2 < t_3$.

2.3. Parameterizations

`PAUP`* (Swofford, 2002) is a popular software which provides the maximum likelihood estimate of an ultrametric tree based on the likelihood function in Eq (2.2). With $\mathbf{t}$, the optimization of Eq (2.2) would be performed with the inequality constraint $t_1 < t_2 < t_3$. `PAUP`* implemented three simple parameterizations of node ages of an ultrametric tree: `Thorne`, `Rambaut`, and `relAge`. Hereinafter, the parameterizations are explained using the ultrametric tree $(T, \mathbf{t})$ in Figure 2.1, where $\mathbf{t} = (t_1, t_2, t_3)$ is the vector of the ages of internal nodes.

The `Thorne` parameterization uses (d_1, d_2, d_3), where $d_3 = t_3$ is the root age, and $d_i = \dfrac{t_{i+1} - t_i}{t_{i+1}}$, for $i = 1, 2$, is the ratio of the age difference between an internal node and its parent to the parent's age. The bound on the root age can be enforced by log transformation, while the ratio parameters are constrained to the interval $[0, 1]$.

The `Rambaut` parameterization uses (c_1, c_2, c_3), where c_i is the age difference between an internal node and its oldest child node, that is, $c_1 = t_1, c_2 = t_2$, and $c_3 = t_3 - t_2$ in Figure 2.1(b). In the optimization procedure, c_i's are constrained to be in $[0, L]$, where L is a preset maximum length of a branch.

The `relAge` parameterization uses (r_1, r_2, r_3), where $r_3 = t_3$, and $r_i = \dfrac{t_{i+1}}{t_i}$ for $i = 1, 2$. For $i = 1, 2$, r_i is the ratio of the age of the parent node of an internal node to the node itself. The bound on the root-age can also be enforced by log transformation. The node-age ratios must be held within $[1, \infty)$, but for numerical stability and convergence rate, the upper bound is preset to a large number L'.

Each of the parameter vectors for the three parameterizations is a one-to-one correspondence with $\mathbf{t} = (t_1, t_2, t_3)$, where $0 < t_1 < t_2 < t_3$. With any of the three parameterizations, the optimization is not constrained with inequality conditions. In principle, the choice of parameterizations does not affect the inference of an ultrametric tree. However, whether or not some parameters are close to bounds depends on parameterizations. For example, if a

tip is very short as $t_1 \approx 0$ in Figure 2.1(a), then $d_1 \approx 1$ (`Thorne`), $c_1 \approx 0$ (`Rambaut`), and $r_1 \approx L'$. Another example is the case for which the ages of two internal nodes are very similar to each other, such as $t_1 \approx t_2$ in Figure 2.1(a). Then $d_1 \approx 0$ (`Thorne`), $c_1 \approx c_2$ (`Rambaut`), and $r_1 \approx 1$ (`relAge`).

It is important to examine the reliability of parameterizations in the inference method, because estimated ultrametric trees or parts of them such as node ages, ranked topologies, or topologies can be subsequently analyzed. For example, when the tree branch is very short, it is not difficult to observe zero or even negative branch lengths estimated by tree estimation software. In such cases, the trees with zero or negative branch lengths are removed from the next analysis, or the branch lengths are manually replaced by a small number which is likely to be larger than the true value. Therefore, in those cases, the choice of parameterizations may result in different performances.

3. Simulation study

The effect of parameterizations of branch lengths on the performance of `PAUP*` was examined. We considered phylogenetic trees with the fixed ranked topology (Figure 2.1 (b)) and diverse branch lengths. First, t_1 was varied as $t_1 = 0.01, 0.001$, and 0.0001, while $t_2 = 0.02$ and $t_3 = 0.03$ were fixed. Secondly, t_2 was varied as $t_2 = 0.02$, 0.011, 0.0101, while $t_1 = 0.01$ and $t_3 = 0.03$ were fixed. Therefore, the differences between t_1 and t_2 varied as $0.01, 0.001$, and 0.0001. We used `seq-gen` (Rambaut and Grassly, 1997) to simulate 100 datasets of DNA sequences of length 1,000 from Eq (2.2), given each tree and the JC model (Eq. 2.1). Each simulated dataset was analyzed to estimate an ultrametric tree with the three parameterizations of `PAUP*`. The JC model was assumed, and `PAUP*` defaults were used in all cases with exception of the branch parameterization.

Estimation of very short tips

DNA sequences were generated under phylogenetic trees (Figure 2.1) with fixed $t_2 = 0.02$ and $t_3 = 0.03$ and with each of varied $t_1 = 0.01, 0.001, 0.0001$. In each case, the correct ranked topology was estimated, and it was not affected by the branch parameterizations. When $t_1 = 0.01$, the average of $\hat{t}_1$ was around 0.01 with standard error (s.e.) 0.0025 (Table 3.1). Other branch lengths such as t_2 and t_3 were correctly estimated by the three branch parameterizations. In the case of `relAge` parameterization, the average $\hat{t}_1$ value was around 0.001, which is close to the true value of $t_1 = 0.001$. However, t_1 was overestimated if the true value is as small as 0.0001. When `Thorne` and `Rambaut` parameterizations were considered, $\hat{t}_1$ was zero in 14% of 100 replicates when $t_1 = 0.001$, and 63% and 56%, respectively, when $t_1 = 0.0001$ (Table 3.3). Disregarding the cases with $\hat{t}_1 = 0$, t_1 was also overestimated by `Thorne` and `Rambaut` parameterizations (Table 3.2).

In principle, the branch length parameterizations do not affect the estimations of branch lengths. However, estimations of very small branch lengths tends to be biased. Given the same precision, `Thorne` and `Rambaut` parameterizations can result in branch lengths of zero more often than `relAge`. In other words, the parameters corresponding to t_1 were often estimated at the boundaries: $\hat{d}_1 = 1$ (`Thorne`) and $\hat{c}_1 = 0$ (`Rambaut`). If estimated trees including zero branch lengths are eliminated from the next step of analysis, the `Thorne` and `Rambaut` parameterizations could eventually affect the final result of the analysis and have smaller power.

Table 3.1 Average of estimated values $\hat{t}_1$ from 100 replicates. The standard errors are in parentheses.

t_1	`relAge`	`Thorne`	`Rambaut`
0.01	0.010062 (0.002505)	0.010062 (0.002505)	0.010062 (0.002505)
0.001	0.001032 (0.000760)	0.000990 (0.000807)	0.000990 (0.000807)
0.0001	0.000357 (0.000141)	0.000120 (0.000247)	0.000120 (0.000247)

Table 3.2 Average of estimated values $\hat{t}_1$ from the replicates after disregarding the cases of $\hat{t}_1 = 0$. The standard errors are in parentheses.

t_1	`Thorne`	`Rambaut`
0.001	0.001151 (0.000755)	0.001151 (0.000755)
0.0001	0.000324 (0.000316)	0.000273 (0.000313)

Table 3.3 Percentage of $\hat{t}_1 = 0$ from 100 replicates. Note that the same true values of $t_2 = 0.02$ and $t_3 = 0.03$ are assumed.

t_1	`relAge`	`Thorne`	`Rambaut`
0.01	0	0	0
0.001	0	14	14
0.0001	0	63	56

Effect of very similar node ages

While $t_1 = 0.01$ and $t_3 = 0.03$ of the ultrametric tree in Figure 2.1(b) is fixed, different values of $t_2 = 0.02$, 0.011, and 0.0101 were assumed. In these cases, the age difference between two internal nodes t_2 and t_1 are small, specifically $t_2 - t_1 = 0.01$, 0.001, and 0.0001. If the ages of two internal nodes are very similar to each other, such as $t_2 - t_1 \approx 0$ in Figure 2.1(b), then the values of parameters by `Thorne` and `relAge` parameterization are close to the boundaries: $d_1 \approx 0$ (`Thorne`), $c_1 \approx c_2$ (`Rambaut`), and $r_1 \approx 1$ (`relAge`).

As the age difference between two internal nodes became smaller, more ranked topologies were incorrectly estimated, as shown in Figure 2.1(c) (Table 3.4). If $t_1 = 0.01, t_2 = 0.02$, and $t_3 = 0.03$, then the true ranked topology was always estimated by the three parameterizations. If t_2 was reduced to $t_2 = 0.011$ ($t_2 - t_1 = 0.001$), then 61% of the estimated ranked topologies were correct with `relAge` and `Thorne`, and 65% with `Rambaut`. If $t_2 = 0.0101$ ($t_2 - t_1 = 0.0001$), 49% were correct with `relAge` and `Thorne`, and 54% with `Rambaut`. Because the two node ages themselves, $c_1 = t_1$ and $c_2 = t_2$, are estimated with `Rambaut`, the accuracy of ranked topology estimation is better than that of `relAge` and `Thorne` which use ratios or differences of the node ages. Therefore, if estimated ranked topologies are considered in subsequent analyses, `Rambaut` has less influence on the final result as compared to other parameterizations.

The choice of three parameterizations did not affect the estimation of t_2. Given that the correct ranked topology was estimated by `Rambaut`, the averages of $\hat{t}_2$ were 0.020262, 0.011714 (s.e. 0.0018), and 0.012004 (s.e. 0.00203) for $t_2 = 0.02, 0.011$, and 0.0101, respectively. Both `relAge` and `Thorne` resulted in similar estimates. When the ranked topology was incorrectly estimated as in Figure 2.1(c), $\hat{t}_1$ and $\hat{t}_2$ were close to each other and close to their true values, respectively.

Table 3.4 Percentage of cases in which the true ranked topology was estimated. Note that the true values of $t_1 = 0.01$ and $t_3 = 0.03$ are fixed in the simulation.

t_2	`relAge`	`Thorne`	`Rambaut`
0.02	100	100	100
0.011	61	61	65
0.0101	49	49	54

4. Real data analysis

To compare the performance of three parameterizations on real data analysis, we analyzed 1,000 alignments of six DNA sequences from chimpanzees. The data was previously analyzed to estimate the population/species-level evolutionary history of two common chimpanzee subspecies, *Pan troglodytes (P. t.) troglodytes* and *P. t. verus* (Chung and Hey, 2017), but not the phylogenetic tree of an alignment. The alignments are located on chromosomes distantly enough to assume their independence (Chung and Hey, 2017). Moreover, alignments can have different phylogenetic trees because of biological processes including recombination. Each alignment was analyzed using each of the three parameterizations by `PAUP`*.

Table 4.1 compares the branch lengths of the ultrametric trees estimated by the three parameterizations. A phylogenetic tree with six tips has six terminal branches directly connecting tips and four internal branches connecting between internal nodes. When the `relAge` parameterization was considered, 262 out of 1,000 ultrametric trees had one or more terminal or internal branches of length zero. Compared to `relAge`, `Thorne` and `Rambaut` estimated much more trees with zero-length branches: 810 trees by `Thorne` and 865 trees by `Rambaut` (Table 4.1). Moreover, around 70% of the 1,000 estimated trees by the `Thorne` and `Rambaut` parameterizations had one or more terminal branches of length zero, while around 5% were in such trees by `relAge`. This result is consistent with the simulation result, which is that `Thorne` and `Rambaut` parameterizations can result in zero-length terminal branches more often than `relAge`. The `relAge` parameterization resulted in 216 trees with zero-length internal branches, which are less than a half of such trees by `Thorne` and `Rambaut` (510 and 612 trees, respectively). Therefore, if these estimated ultrametric trees with non-zero length branches are used to a subsequence analysis, the `relAge` parameterization, rather than `Thorne` and `Rambaut`, could provide more information in terms of the number of trees.

Table 4.1 Number of the ultrametric trees with one or more branches of length zero, out of 1,000 trees estimated from DNA alignments of chimpanzees.

Branches of length zero	`relAge`	`Thorne`	`Rambaut`
terminal branches	51	701	718
internal branches	217	510	621
any branches	262	810	865

We also compared the three parameterizations in terms of the topologies of the 1,000 ultrametric tree estimations. We note that there are 945 possible rooted topologies on six terminal nodes (Semple and Steel, 2003). Since the true ultrametric trees of the DNA alignments from chimpanzees are unknown and can be different from each other, it is difficult to evaluate the accuracy of the three parameterizations. Nonetheless, we compared the topologies estimated by the three parameterizations for each alignment and computed the number of cases that

Table 4.2 Number of the cases that all, any two, or none of the three parameterizations resulted in the same ultrametric tree topology estimation.

Methods	No. of cases	
three parameterizations	434	
two parameterizations	357	
`relAge` and `Thorne`		156
`relAge` and `Rambaut`		101
`Thorne` and `Rambaut`		100
none	209	

all, any two, or none of the three parameterizations resulted in the same topology estimation (Table 4.2). The three parameterizations estimated the same topology in 434 cases out of 1,000 alignments, while different topologies were estimated by the three parameterizations in 209 cases. In the remaining 357 cases, the estimated topologies were the same by two of the three parameterizations. Among three pairs of the three parameterizations, `relAge` and `Thorne` resulted in the same topologies around 1.5 times more often than other pairs (Table 4.2). An important caveat is thus that the majority rule cannot be applied to choose the best parameterization. Through the simulation, we found the case that `Rambaut` was more accurate than `relAge` and `Thorne` (Table 3.4), although the estimated ranked topologies by `relAge` and `Thorne` were always the same in the simulation.

5. Conclusion

In this study, the performance of `PAUP`* software was evaluated to estimate an ultrametric phylogenetic tree from DNA sequences of a group of organisms. A simulation study was conducted to examine the effect of the three parameterizations of branch lengths. In particular, we consider two cases in which branch lengths of zero or wrong ranked topologies were frequently estimated. From the simulation, `relAge` did not provide the case of zero branch lengths, but very short tips can be overestimated by the three parameterizations. When a tree has internal nodes with similar ages, `Rambaut` parameterization resulted in more accurate ranked topologies than others. In general, however, the order of the internal nodes with very similar ages tended to be randomly determined by all the methods.

The cases with short terminal or internal branches can be more easily observed when a higher number of organisms are collected for analysis. We analyzed 1,000 alignments of six DNA squences from chimpanzees using the three parameterizations by `PAUP`*. The analyses of simulated and real data had a steady result that `Thorne` and `Rambaut` parameterizations can result in branch lengths of zero more often than `relAge`. When we compared the topoologies, there were more than 50% cases that different topologies were estimated by two or all of the three parameterizations. When a topology is of main interest, it is difficult to choose the best parameterization. Thus, a further simulation study is needed to examine the performance of the topology estimation on six or more sequences by the three parameterizations.

When an analysis is a procedure of several steps and estimated ultrametric trees are analyzed in subsequent steps, incorrectly estimated branch lengths or tree structures such as ranked topology or topology can affect the final result of the analyses (Gavryushkin and Drummond, 2016). However, there is no best parameterization that always outperforms

others. Therefore, the estimated trees require careful examination to check for the presence of zero branch lengths or internal nodes with similar ages. If necessary, the analysis result with different parameterizations should be compared to determine the way of parameterizations.

In subsequent analysis, the effect of using estimated trees can be reduced by incorporating the uncertainty of tree estimations. One method is to add uncertainties such as standard errors to the uncertainty of the final result. However, the problem persists, particularly with the uncertainty of discrete tree component such as ranked topologies, because parameterizations of branch lengths affect the space properties of a metric of ultrametric trees (Gavryushkin and Drummond, 2016). Therefore, further studies are necessary to investigate the effect of parameterizations on uncertainty estimations of continuous and discrete components of trees. An alternative method to incorporate the uncertainty of tree estimation is to use the posterior distribution of ultrametric trees as a framework of Bayesian inference (Ané *et al*., 2007; Chung and Hey, 2017). However, computational complexity is one of the major barriers to using the posterior distribution of trees for population/species-level inference (Chung, 2019).

References

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. and Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, **3**, 407-415.

Ané, C., Larget, B., Baum, D. A., Smith, S. D. and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*, **24**, 412-426.

Chung, Y. (2019). Recent advances in Bayesian inference of isolation-with-migration models. *Genomics & Informatics*, **17**, e37.

Chung, Y. and Hey, J. (2017). Bayesian analysis of evolutionary divergence with genomic data under diverse demographic models. *Molecular Biology and Evolution*, **34**, 1517-1528.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368-376.

Felsenstein, J. (2004). *Inferring phylogenies*, Sinauer Associates, Inc., Sunderland, MA.

Garrison, N. L., Rodriguez, J., Agnarsson, I., Coddington, J. A., Griswold, C. E., Hamilton, C. A., Hedin, M., Kocot, K. M., Ledford, J. M. and Bond, J. E. (2016). Spider phylogenomics: Untangling the spider tree of life. *PeerJ*, **4**, e1719.

Gavryushkin, A. and Drummond, A. J. (2016). The space of ultrametric phylogenetic trees. *Journal of Theoretical Biology*, **403**, 197-208.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160-174.

Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L., and Kimball, R. T. (2015). Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, **33**, 1110-1125.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754-755.

Jeon, Y. and Cho, H. (2019). Model based hybrid decision tree. *Journal of the Korean Data & Information Science Society*, **30**, 515-524.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of protein molecules*, Academy Press.

Lauritzen, S. L. (1996). *Graphical models*, Oxford University Press.

Lee, M. S. and Ho, S. Y. (2016). Molecular clocks. *Current Biology*, **26**, R399-R402.

Nascimento, F., Reis, M. and Yang, Z. (2017). A biologist's guide to bayesian phylogenetic analysis. *Nature Ecology and Evolution*, **1**, 1446-1454.

Park, G. (2019) Discovering a fine dust pathway via directed acyclic graphical models. *Journal of the Korean Data & Information Science Society*, **30**, 67-76.

Rambaut, A. and Grassly, N. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, **13**, 235-238.

Rogers, J. S. (2001). Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Systematic Biology*, **50**, 713-722.

RoyChoudhury, A., Willis, A. and Bunge, J. (2015). Consistency of a phylogenetic tree maximum likelihood estimator. *Journal of Statistical Planning and Inference*, **161**, 73-80.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406-425.

Semple, C. and Steel, M. (2003). *Phylogenetics*, Oxford university press, New York, NY.

Siepel, A. (2009). Phylogenomics of primates and their ancestral populations. *Genome Research*, **19**, 1929-1941.

Swofford, D. (2002). *PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0*, Sinauer Associates.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, **17**, 57-86.

Truszkowski, J. and Goldman, N. (2016). Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. *Systematic Biology*, **65**, 328-333.

Yang, Z. (1995). Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. *Journal of Molecular Evolution*, **40**, 689-697.

Yang, Z. (2006). *Computational molecular evolution*, Oxford university press.