Mathematical Statistics
Stockholm University

Phylogenetics and inference

Fredrik Olsson

Examensarbete 2009:6

Postal address:
Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:
http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2009:6**,
http://www.math.su.se/matstat

Phylogenetics and inference

Fredrik Olsson*

May 2009

Abstract

A phylogenetic tree describes the relatedness between species in an evolutionary context. A phylogenetic tree can be reconstructed or estimated by comparing DNA-sequences for a number of species. In this thesis we are simulating trees according to a linear birth-death process. The DNA-evolution in the trees are simulated according to the Jukes-Cantor model. By performing the analyses with a Bayesian method and a maximum likelihood method, we study how the inference change when increasing the number of taxa and/or the length of the sequences. We compare the estimates with the simulated tree and use three different measurements. For our comparison of the two methods we are using the programs MrBayes and PHYLIP. The main result is that the estimates becomes better when increasing the length of the sequence but not when we increase the number of taxa. We also observe that PHYLIP, which uses a maximum likelihood method, has a tendency to perform better than MrBayes, which use a Bayesian method.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: fredriko@math.su.se. Supervisor: Tom Britton.

Acknowledgments

I want to thank my supervisor, Professor Tom Britton, for his encouraging and enthusiastic support with this thesis. I also want to thank Sebastian Höhna for his suggestions and feedback.

Contents

1 Introduction

Phylogenetics is the study of how different species are related to each other. We assume that a group of species has evolved from a common ancestor and we can illustrate their relatedness in a phylogenetic tree. In a phylogenetic tree the root is the common ancestor and the tips are all the extant species.

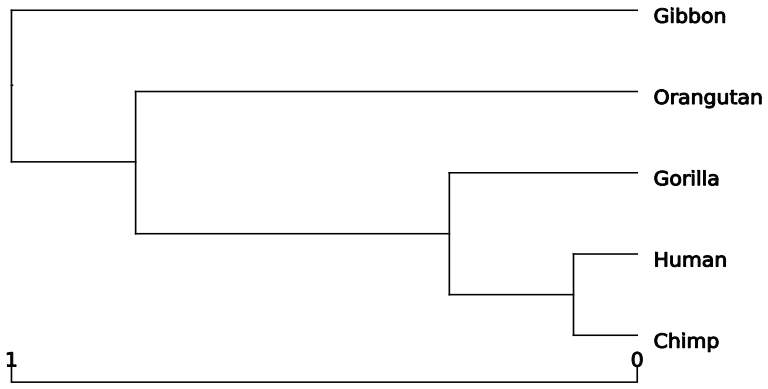

Figure 1: A phylogenetic tree for the great and lesser apes

Let us say that we have a group of currently living species and want to reconstruct their phylogenetic tree. One way to do that is to look at the species DNA-sequences (DeoxyriboNucleic Acid). If we collect DNA-sequences from the species and observe how much they differ we can estimate how close the species are in the evolutionary process.

The positions in a DNA-sequences, further referred to as sites, consist of one out of four nucleotides; A, G, C or T. The DNA-evolution occurs in different ways, however in this thesis we will only consider nucleotide substitutions, when a site changes from one nucleotide to another. We will also assume that all sites evolve independently of each other and that there is a molecular clock. With the molecular clock assumption we mean that all sites in all species DNA-sequences have the same substitution rate for all time periods.

In the evolutionary process a species is eventually divided into two groups, e.g. when a seed from a plant on one side of a mountain travels to and grows on the other side of the mountain. When this event happens the species splits into two new species and they continue to evolve independently of each other.

In this thesis we will simulate phylogenetic trees and DNA-evolution in those trees. We are, with the DNA-sequences at the tips of a tree, going to reconstruct or estimate the phylogenetic tree with two different methods. By comparing the estimates from the two methods with the true tree we can study how the inference depends on the number of taxa and length of the

sequences. The analyses will be done by two computer packages (PHYLIP [3] and MrBayes [6]) and we will compare their performance under different scenarios.

2 Simulating the data

To compare the estimates of the two program packages we need data to analyse. In this thesis we will use simulated data. Simulated data is often used when comparing different inference methods in phylogenetics [7]. One advantage with simulated data compared to real data is that we know the answer and we can see how much the estimated trees differ from the real (simulated) tree. To simulate the data we start by simulating a phylogenetic tree which connects the species and describes there relatedness. After we obtained our tree we simulate DNA-evolution in that tree. Finally, our data will be in the form of a $k \times n$ matrix where k is the number of extant species, for which we want to estimate the phylogenetic tree, and n the length of their DNA-sequences. All simulations were performed in Octave [1] which is a free and open source computer program with many similarities to Matlab.

2.1 Simulating the tree

To simulate a tree we use a linear birth and death process [9]. This process depends on two parameters, the birth rate (λ) and the death rate (μ). The time between two events is exponentially distributed with parameter $i(\lambda+\mu)$, where i is the number of living species in the process. Hence we have that the expected time between two events

$$E[\text{Time until next event}] = \frac{1}{i(\lambda+\mu)}. \tag{1}$$

An event is either a birth or a death, with the probability for a birth

$$P(Birth) = \frac{\lambda}{\lambda+\mu} \tag{2}$$

and the probability for a death

$$P(Death) = \frac{\mu}{\lambda+\mu}. \tag{3}$$

Using the above expressions we can simulate a phylogenetic tree by starting with two branches that evolve from the root and then use the following algorithm:

- Simulate the time until the next event occurs
- Add the simulated time to all active branches where an active branch is a branch that has not been exposed to a death event
- Decide on which branch the event will occur with all branches having equal probabilities to be chosen
- Decide if the event is a birth or a death.
  - If the event is a birth the branch splits into two new branches
  - If the event is a death the branch stops growing
- Repeat the algorithm until we have the desired number of living species in the tree. If a death event occurs at the last branch, restart the process with two new branches
- When we have got the desired number of species we simulate the time until next event and add half of that time to all active branches

After the simulation we erase all the dead branches, which represent extinct species, and then normalize the time in the tree so that the total time from the root to all taxa is 1. This normalization is done because we want all simulated trees to have the same height so that we are able to compare them to each other.

The time in a phylogenetic tree could be expressed in different units, e.g. expected number of substitutions per site or real time. If we are only using DNA-sequences it is impossible to estimate the real time in the tree, for that we need additional information, for example fossils.

In Figure 2 we see an example of a simulated tree where some of the species have gone extinct. However, it is the tree where the extinct species are removed, as in Figure 3, that we use for DNA-simulation and try to estimate with the simulated sequences.

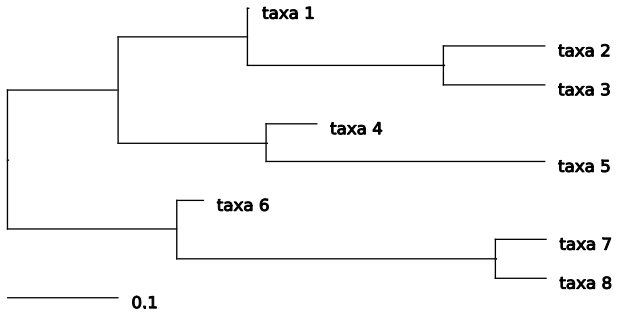

Figure 2: A simulated phylogenetic tree with extinct species

Figure 3: The same phylogenetic tree as in Figure 2 but where all the extinct species have been erased

2.2 DNA-simulation

When simulating the DNA-evolution in a given phylogenetic tree, we start by simulating the sequence for the root. We randomly draw the state for each site with equal probability for all the nucleotides. This sequence then evolves along the branches according to a substitution model until it reaches the tips of the tree.

The substitution model we use is the Jukes-Cantor model [5] which is a continuous time Markov chain. In this model we assume that all sites evolve independently. When a substitution occurs the transition probabilities between all nucleotides are the same. The model does only depend on one parameter, the mean substitution rate, which is the same for all sites.

The continuous time Markov chain has four states, the nucleotides, and by using the Jukes-Cantor model the generator matrix is

$$A = \begin{pmatrix} \frac{-3\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{-3\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{-3\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & \frac{-3\mu}{4} \end{pmatrix} \tag{4}$$

where μ is the mean substitution rate.

The transition matrix, $P(t)$, consist of the probabilities that a site changes from one nucleotide to another at time t, where t is in this case the branch length. We obtain the transition matrix with the forward or backward equa-

tions which have the formal solution

$$P(t) = e^{At} = \sum_{n=0}^{\infty} \frac{t^n A^n}{n!} \tag{5}$$

if the number of states in the chain are finite [9]. In our case the transition matrix is a 4×4 matrix where $P_{ij}(t)$ is the probability that a site changes from nucleotide i to j in t time units.

However, in this case we are able to calculate the transition matrix and express it in a matrix form as

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}. \tag{6}$$

Using the transition matrix we can simulate DNA-evolution for every site by randomize if a site has changed and in the case of a change, randomize to which nucleotide it will change. This way of simulating the DNA-evolution consumes a lot of computer time. Instead we use the assumption that all sites evolve independently and at the same rate.

The number of substitutions along a branch with length t will then be binomial distributed with n equal to the length of the sequence and with the probability that a site has changed in t time units

$$p(t) = \sum_{i=2}^{4} P_{1i} = \frac{3}{4} - \frac{3}{4}e^{-t\mu}. \tag{7}$$

Then, for every branch we can randomly draw the number of substitutions that occurs, at which sites the changes will happen and to which nucleotides the sites will change.

At the end of the simulation we obtain the DNA-sequence for every node and tip in the tree. We pick the sequences from the k extant species and construct a $k \times n$ matrix where the k:th row consists of the DNA-sequence for the k:th species. In Figure 4 we have an example of a small dataset for the species in Figure 3. We can for instance see that the sequences of taxa 2 and taxa 3 are more similar than the sequences of taxa 2 and taxa 8.

```
taxa_2 TGCAAACTCTTAAATAGATGCGTTCGCTATATTATGTTCGTAGAATTCAT
taxa_3 TGCAAACTTTTAAATAGATGCGTTCGCTATATTATGTTCGTAGAATTCAT
taxa_5 TGGAAACTTTTAAATAGATGCGTTCGCTATATTATATTCGTAGAATTCCT
taxa_7 TGGGAACTTTTACATAGATGCGTTCGCTAAATTATATTCGTAGAATTTAT
taxa_8 TGGGAACTGTTACATAGATGCGTTCGCTAAATTATATTCGTAGAATTTAT
```

Figure 4: An example dataset for the species in Figure 3.

2.3 Model parameters

In our simulations we need to choose the birth and death rates which in nature differs between different types of species. Since we normalize the tree height after the simulation we are only interested in the ratio between the two rates. We want the ratio $\frac{\mu}{\lambda}$ to be less than 1, otherwise we have a very small probability that the tree grows large. On the other hand, if the ratio is 0, we get a pure birth process. The shape of the tree also differs depending on the ratio. The splits in a tree simulated with a high ratio tends to be closer to the tips than the splits in a tree simulated with a low ratio [4].

In our study we have used three different values of the ratio, 0, 0.5 and 0.75, and simulated three trees with 100 extant taxa each (see appendix A). In our simulated trees we can see that when the ratio was equal to 0.75 (Figure 19) the splits seems to be closer to the tips than when the ratio equals 0.5 (Figure 18) or 0 (Figure 17).

For each root in the trees we have simulated a DNA-sequence with length 20,000. That sequence has evolved in the tree according to our substitution model. We chose the mean substitution rate so that the expected number of changes per site and time unit was equal to 0.01 in the substitution model.

2.4 Choosing trees and DNA-sequences

From each of the simulated trees with 100 taxa we picked subtrees to analyse. First, we construct a subtree that consists of 3 randomly chosen taxa with the restriction that the subtree should contain the root of the original tree. Then we add randomly chosen taxa to the tree until we get subtrees containing of 5, 10, 20, and 50 species. This procedure is repeated twice so that we finally have 3 subtrees of each 3, 5, 10, 20 and 50 taxa. If we do not have the restriction that the original root should appear in the subtree, there is a large risk that our subtrees have different heights. If the subtrees have different heights it would not be fair to compare them, because they would not have the same amount of time for their DNA-sequences to evolve.

For each subtree we pick a subsequence of 1000 sites to which we add sites until we have subsequences of size 1000, 2000, 5000 and 10000. For every combination of the number of species, k, and sequence length, n, we can form a $k \times n$ matrix which is the data for our analyses.

3 Computer packages

There are many computer packages which simulate, analyse and summarize phylogenetic data, most of them are still developing and implementing new methods. In our thesis we will use two different computer packages to analyse the simulated DNA-sequences, PHYLIP (The PHYLogeny Inference Package) [3] and MrBayes [6].

3.1 PHYLIP

We have used PHYLIP version 3.67 which is a free and open source computer package developed by Joe Felsenstein [3]. In our analyses we used the program Dnamlk which is a part of PHYLIP. Dnamlk analyses DNA-sequences with a maximum likelihood approach and assumes a molecular clock.

3.1.1 Maximum likelihood inference

In maximum likelihood inference we collect data X, in our case the DNA-sequences, and formulate the likelihood function $L(\tau, b(\tau)|X)$ where τ is the tree topology and b(τ) branch lengths for topology τ. The likelihood function expresses how likely the values on the parameters are given the observed data. The values of τ and $b(\tau)$ that maximize the likelihood function will be our estimates for the parameters. Since all the sites are mutually independent this likelihood could, for a given topology, be expressed as a product $\Pi_{i=1}^{n} L(\tau, b(\tau)|X_i)$ where X_i denotes the i:th column in our data matrix.

One problem with maximizing the likelihood is when the number of taxa increases, because then the number of possible topologies grows very fast. It has been shown [2] that the number of rooted bifurcating trees with n taxa is $\frac{(2n-3)!}{(n-2)!2^{n-3}}$. The time to calculate the branch lengths that maximizes the likelihood function for every possible topology would increase fast when the number of taxa in our analyses increases.

To overcome this problem we start with two taxa and successively add taxa to that tree instead of considering all possible topologies. We place every added taxa where it maximizes the likelihood function. When the last taxa is added to the tree it is possible to rearrange the placements of the taxa.

When rearranging, we change the placement for the taxa one by one in order to find a new topology which yields a higher value for the likelihood function [2].

3.2 MrBayes

MrBayes is developed by Fredrik Ronquist, John P. Huelsenbeck and Paul van der Mark, and like PHYLIP it is an open source and free computer program. MrBayes uses a Bayesian approach and analyses the DNA-sequences with the Markov Chain Monte Carlo (MCMC) method.

3.2.1 Bayesian inference

In Bayesian inference we regard the parameters as random variables and we are interested in the parameters posterior distribution. This is the distribution of the parameters given the data and our prior knowledge about the parameters. To find the posterior distribution we use Bayes' theorem to combine the parameters prior distributions, which describes our prior beliefs or knowledge about the parameters, with the observed data. If we do not have any beliefs or knowledge about the parameters we try to construct a non-informative prior.

The posterior distribution is often difficult to calculate especially for complex problems like the ones in phylogenetics. But in the end of the 20th century the Bayesian approach became popular for phylogenetics when a new method, MCMC, was proposed to solve the posterior problem [8]. In MCMC we formulate a Markov chain with the posterior distribution as its stationary distribution. If we run the Markov chain until it approximately reaches it stationary distribution and then take samples from it, we get approximate samples from the unknown posterior distribution.

4 Analysis

Both programs have a lot of different options and we will explain our choices in the following section. Our goal is to use the same models in both programs to get a fair comparison of them.

4.1 PHYLIP

The analyses with PHYLIP were performed with the program DNAMLK. The substitution model we used is the Jukes-Cantor model and we enabled

the global rearrangements option. With this option we try to rearrange the topology after the last taxa is added as we described earlier in section 3.1.1.

4.2 MrBayes

In MrBayes there are a number of parameters and we need to specify their prior distributions before we can run the analyses. For the branch length we used a birth and death model with a molecular clock. We set the prior distribution for the birth rate and the death rate to be uniformly distributed between 0 and 10, which is default in MrBayes. We used the Jukes-Cantor model as our substitution model and for the remaining parameters we used the default prior distributions of MrBayes.

A problem when using MrBayes is knowing when to stop the analysis, when is the sampled distribution close enough to the posterior distribution? To decide when to stop we used a stop-rule which is implemented in MrBayes. For every analysis with MrBayes we ran two separate Markov chains and at every 1000th generation MrBayes calculate the average standard deviation of split frequencies between these two runs. This value is calculated from 75 % of the past generations and is a measure on how much the two samples differ. When this value gets low the samples become more and more simular and we consider that we have reached the stationary distribution. In our analysis we used the default stopvalue which is 0.01.

After we decided that we have reached the stationary distribution we summarized the samples with the sump and sumt commands in MrBayes. In the analysis we discarded the first 25% of the generations when we decided that the chains had converged, therefore we omit the first 25% samples when we summarize the analysis and consider that as the burnin period.

4.3 Program performance

The time to run an analysis with the two programs differs, especially when the amount of data increases. The mean time of 10 runs with a data set consisting of 20 taxa with sequence lengths of 5000 sites, on an AMD Turion X2 Ultra Dualcore Mobile ZM-82 2200 MHz 1MB cache, was 326 seconds for MrBayes and 42 seconds for PHYLIP. Hence, for this dataset PHYLIP was almost 8 times faster than MrBayes.

Both MrBayes and PHYLIP also encountered problems when analysing data sets with large number of taxa and short DNA-sequences. PHYLIP produced a segmentation fault and MrBayes froze at the beginning of the analyse. To solve these problems we changed the input order in PHYLIP and reran the analysis with MrBayes. If both programs had a problem to analyse the same

dataset we chose a new sequence to analyse for that tree. However this did not solve the problem for all the analyses when we had 50 taxa with 1000 sites as our data, therefore this data is discarded from the analysis.

5 Comparing two estimated trees

In order to compare two estimated trees with each other we need some measurements on how good the estimates are. A measurement also helps us to study how the inference changes when the number of taxa or the length of the sequences changes. We have used two measurements in this thesis, the proportion of correct splits and the absolute branch length distance. We also study how the estimates of the node-heights, in a tree with three taxa, changes as we add new taxa to that tree.

5.1 Topology measure

In order to measure how close the estimated topology is to the true topology we study the proportion of correct splits. This meassure is similar to the symmetric distance metric presented by Robinson and Foulds[10]. From each split in a topology there are a number of taxa that have descended and the split divides those taxa into two groups. If we compare a split in an estimated tree with the true topology we consider the split to be correct if we, for any split in the true topology, can construct the exact same groups of taxa as we observed for the split in the estimated topology. The proportion of correct splits is then the number of correct splits divided by the total number of splits in the tree.

5.2 Branch length measure

In order to compare branch lengths in trees with different topologies we study the absolute branch length distance between the estimated tree and the simulated one. Let x_{ij} be the true sum of the length of branches which connects the two taxa i and j in a simulated tree with k taxa, and let $\hat{x}_{ij}$ be the corresponding sum in the estimated tree. This distance is then defined by

$$D = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |x_{ij} - \hat{x}_{ij}| \tag{8}$$

and a low value of D means that the estimated branch lengths are close to the simulated tree. With this measures we see how much the estimated tree

differs from the true tree and we do not need to estimate the correct topology, in fact we can compare the branch lengths in estimates with different topologies.

5.3 Difference between the measurements

Which of these measurements is preferable? To answer this question we must know what we are interested in and what we consider to be a good estimate. In Figure 5 we have the true tree and we have two estimated trees in Figure 6 and Figure 7. If we look at the proportion of correct splits we see that all of the splits in Figure 7 are correct but only one third of the splits in Figure 6 are correct. On the other hand, the D-value for the tree in Figure 6 is much lower than it is for the tree in Figure 7.

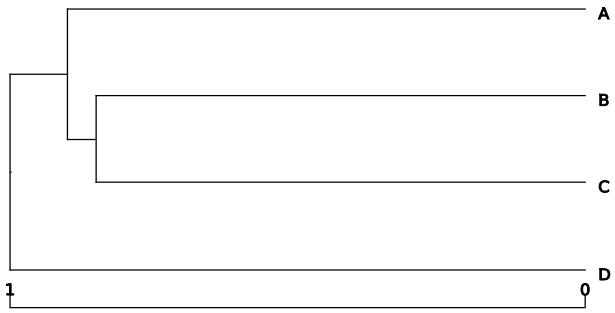

Figure 5: The true topology and branch lengths

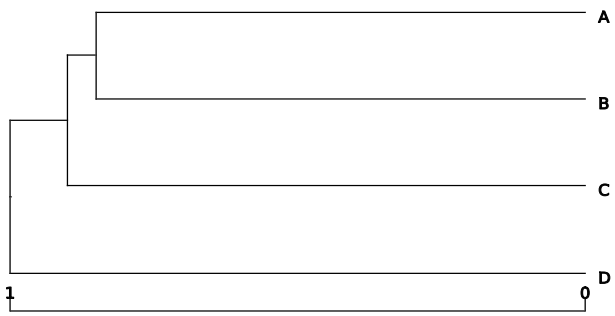

Figure 6: Estimated tree with incorrect topology

5.4 Adding taxa with fixed sequence length

To study the performance of the programs when we hold the length of the sequence fixed and successively add taxa we start by looking at a normalized tree with three taxa. Such a tree has 2 splits where the height of the split,

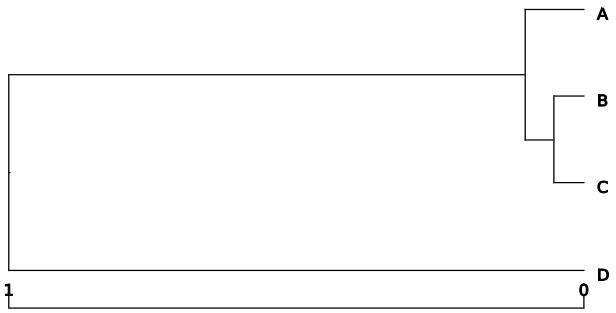

Figure 7: Estimated tree with correct topology

that is not the root, lies between zero and one. By comparing that height with the corresponding height in the true topology we can study how the inference changes when we add new taxa to the tree.

6 Results

6.1 Proportion of correct splits

In Table 1, 2 and 3 we have the mean proportion of correct splits for three trees when we have various number of taxa and length of the sequences. When we analyse a tree with few taxa there are only a few possible values for each cell. Therefore, it is not surprising if cells with few taxa have the same mean proportion for different sequence length. Because we discarded the data sets with 50 taxa and 1000 sites we have no estimate for these cells.

Both program estimated the correct splits for all trees that consisted of three taxa for all considered sequence length. As we can see the proportion seems to increase as we increase the length of the sequences and fix the number of taxa. This is expected because when we increase the length of the sequences we increase the amount of information.

When we hold the sequence length fix and increase the number of taxa we see that the proportion in table 1 and 3 decreases but that pattern does not appear in table 2. One possible reason why both programs did not find many correct splits when we used 5 taxa in Table 2 is that many of the splits in these topologies are close to each other and they are therefore hard to estimate.

In Table 2, the proportion of correct splits decreased dramatically when we analysed 5 taxa and the sequence length increased from 5000 sites to 10000 sites. The splits in these trees were hard to estimate because they were close to each other. It is possible that, by chance, the additional 5000

sites increased the support for a topology separate from the true topology. This explanation becomes more reasonable because both program behaved similar.

In Table 3, the tree was simulated with the ratio $\frac{\mu}{\lambda} = 0.75$ and we estimated more correct splits than in the other two tables. In that tree the splits tend to be near the tips and it should therefore be easier to find the splits, when we have few taxa, than for an equally sized tree but with the splits near the root.

If we compare the programs to each other we see that they estimate approximately the same proportion of correct splits in most cases and we cannot see with this data that one of the programs performs better than the other.

		Length of sequence			
k	Program	1000	2000	5000	10000
3	MrBayes	1	1	1	1
	PHYLIP	1	1	1	1
5	MrBayes	0.83	0.83	0.83	0.83
	PHYLIP	0.83	0.83	0.83	0.83
10	MrBayes	0.56	0.85	0.93	0.93
	PHYLIP	0.63	0.85	0.85	0.85
20	MrBayes	0.67	0.77	0.86	0.79
	PHYLIP	0.79	0.81	0.89	0.82
50	MrBayes	Na	0.73	0.80	0.85
	PHYLIP	Na	0.65	0.86	0.85

Table 1: Every cell consists of the mean proportion of correct splits for three trees. The subtrees are sampled from a tree simulated with $\frac{\mu}{\lambda} = 0$

		Length of sequence			
k	Program	1000	2000	5000	10000
3	MrBayes	1	1	1	1
	PHYLIP	1	1	1	1
5	MrBayes	0.58	0.58	0.83	0.50
	PHYLIP	0.58	0.58	0.83	0.42
10	MrBayes	0.70	0.67	0.70	0.74
	PHYLIP	0.70	0.70	0.78	0.81
20	MrBayes	0.79	0.77	0.82	0.96
	PHYLIP	0.79	0.81	0.84	0.96
50	MrBayes	Na	0.76	0.89	0.92
	PHYLIP	Na	0.66	0.88	0.92

Table 2: Every cell consists of the mean proportion of correct splits for three trees. The subtrees are sampled from a tree simulated with $\frac{\mu}{\lambda} = 0.5$

		Length of sequence			
k	Program	1000	2000	5000	10000
3	MrBayes	1	1	1	1
	PHYLIP	1	1	1	1
5	MrBayes	1	1	1	1
	PHYLIP	1	1	1	1
10	MrBayes	0.85	1	1	1
	PHYLIP	1	1	1	1
20	MrBayes	0.82	0.93	1	1
	PHYLIP	0.74	0.87	1	1
50	MrBayes	Na	0.80	0.90	0.92
	PHYLIP	Na	0.78	0.88	0.90

Table 3: Every cell consists of the mean proportion of correct splits for three trees. The subtrees are sampled from a tree simulated with $\frac{\mu}{\lambda} = 0.75$

6.2 Comparing branch lengths when increasing sequence length

In figures 8-12 we can see the absolute branch length difference for our estimated trees with the number of taxa fixed. The mean D-value decreases as we increase the sequence length in all figures except in one case when we analyse three taxa. The spread among the estimate also decreases as the sequences gets longer which is to expect because we add more data to the analyses. We also notice that the lines, which represent the mean D-value for the estimates from both the programs, seems to follow each other, especially when the sequences becomes longer.

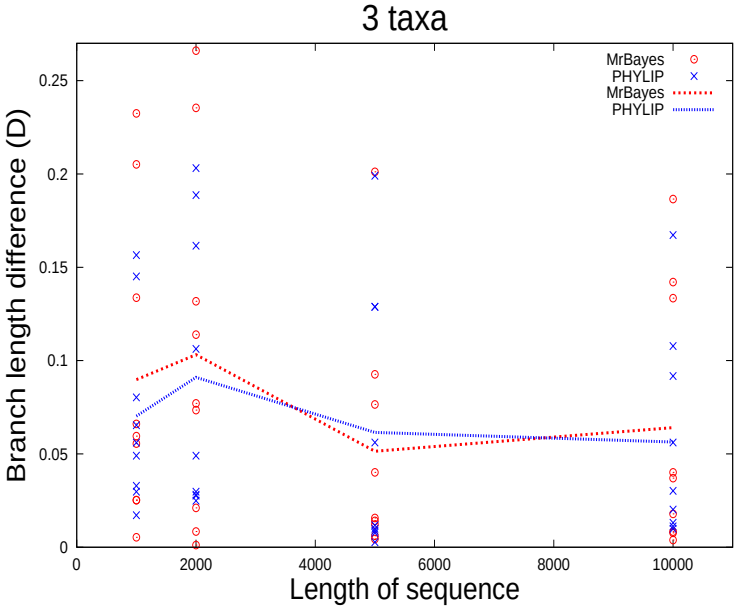

Figure 8: D-values for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every length of the sequence there are 9 observations for each of the programs.

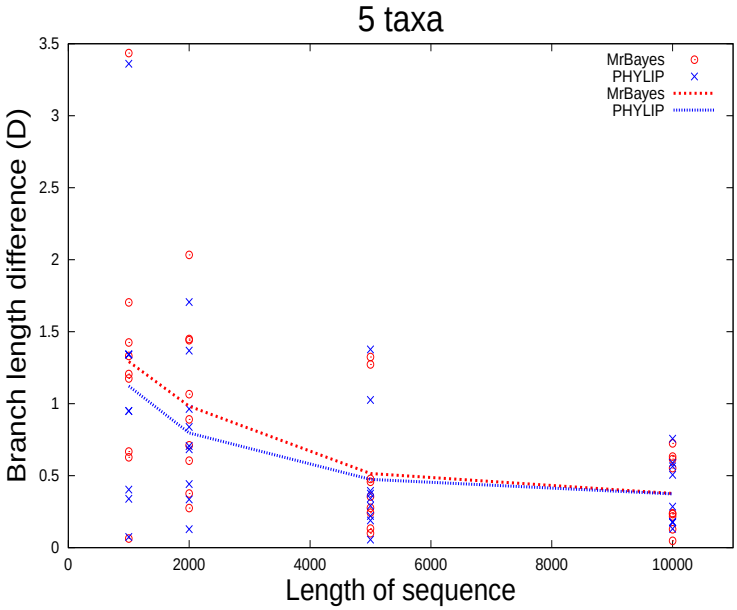

Figure 9: D-values for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every length of the sequence there are 9 observations for each of the programs.

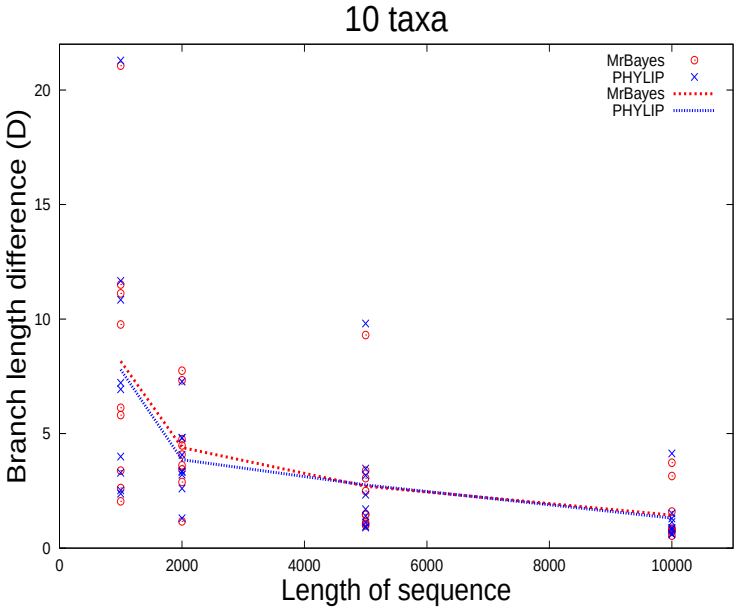

Figure 10: D-values for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every length of the sequence there are 9 observations for each of the programs.

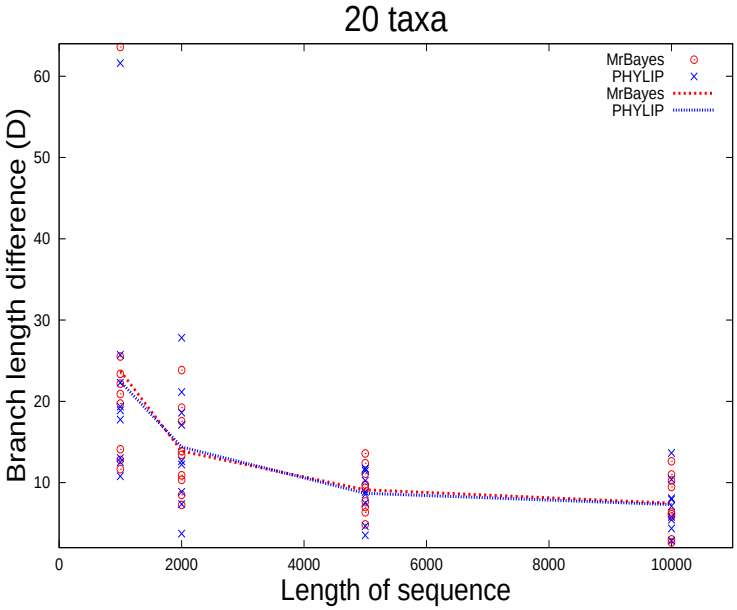

Figure 11: D-values for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every length of the sequence there are 9 observations for each of the programs.

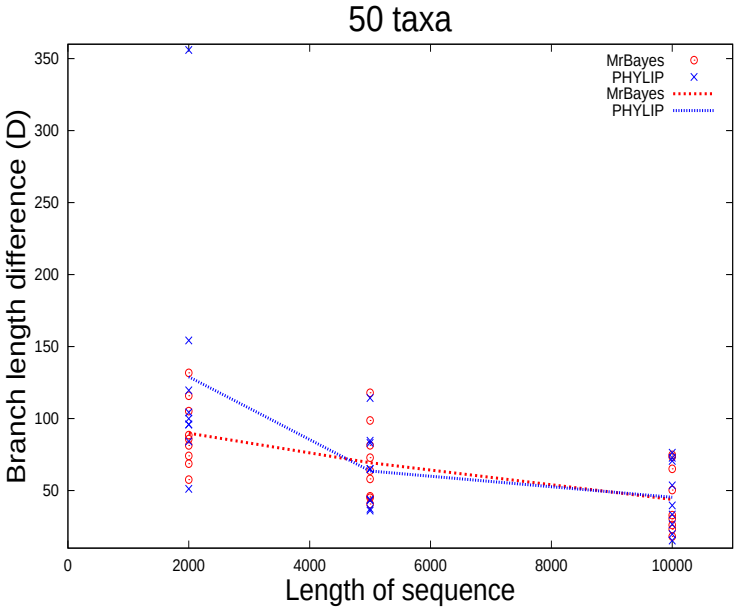

Figure 12: D-values for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every length of the sequence there are 9 observations for each of the programs.

In our analyses every data set is analysed with both programs and we have therefore obtained paired observations which we can compare to see if there are any difference between the programs. By performing a t-test for the differences of the D-values, for every combination of taxa and sequence length, we see in Table 4 that there are two significant differences with a 5% significance level. In both these test did PHYLIP have a significant lower mean D-value than MrBayes.

Because of the number of tests we perform the probability that we would get two significant results by chance is high and we should consider to modify the significance level. Another way of constructing a test with 5% significance level for the overall difference is to construct a new t-statistic from the t-values in table 4 which are standalized and comparable. With that test we get a P-value of 0.06 and we cannot reject the hypothesis that there is a difference between both programs estimates of the total branch-length distance on a 5% significance level. However, this low P-value does indicate a tendency that PHYLIP is performing better with respect to the absolute branch length difference measure.

		Length of sequence			
		1000	2000	5000	10000
3 taxa	T-value	1.24	0.75	-1.44	1.10
	P-value	0.25	0.48	0.19	0.30
5 taxa	T-value	**2.76**	**3.41**	0.94	0.05
	P-value	**0.02**	**0.01**	0.37	0.96
10 taxa	T-value	0.92	1.42	-0.57	0.56
	P-value	0.39	0.19	0.59	0.59
20 taxa	T-value	1.22	-0.49	0.58	0.35
	P-value	0.26	0.64	0.58	0.74
50 taxa	T-value	Na	-1.22	0.65	-1.21
	P-value	Na	0.26	0.53	0.26

Table 4: T-values and P-values when testing if there is any difference of the estimates of the branch lengths

6.3 Comparing branch lengths when adding taxa

If we, for a fixed sequence length, add taxa to our analyses we can see in Figure 13-16 that our estimates in most cases do not improve. The spread decreases a little bit, but in Figure 13 and in Figure 15 the mean value is almost not affected. If we want to estimate the node-height the precision should increase if we add new taxa below the split [11], altought only at a rate proportional to the logarithm of the number of taxa. In our case many of the taxa are added above the split and that could be a reason why we do not see an improvement of the estimates.

If we compare the programs we see that the mean values in Figure 13-16 at many places are lower for MrBayes than for PHYLIP. To test if there is a difference we can use the fact that we have pairwise observations as we did in section 6.2. With that test we do not get any significance difference and by performing a test with overall significance level of 5 % we obtain a P-value of 0.2.

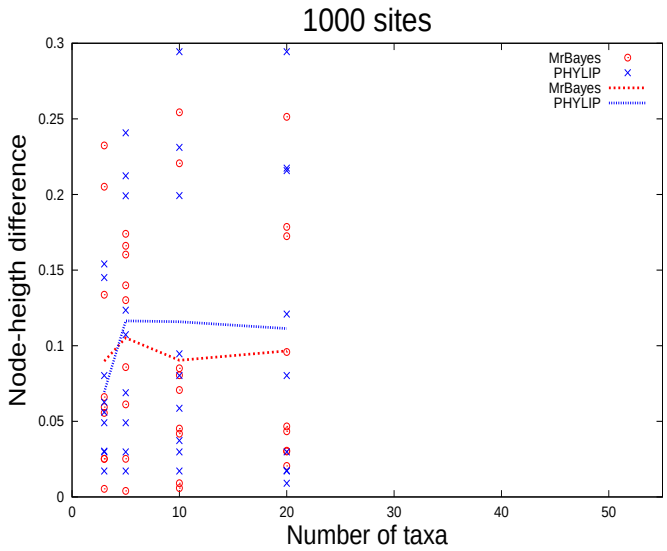

Figure 13: Node-height difference for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every number of taxa there are 9 observations for each of the programs.

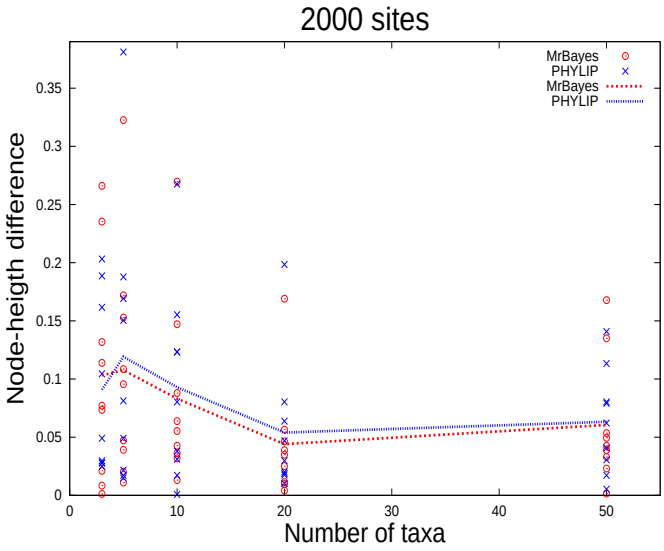

Figure 14: Node-height difference for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every number of taxa there are 9 observations for each of the programs.

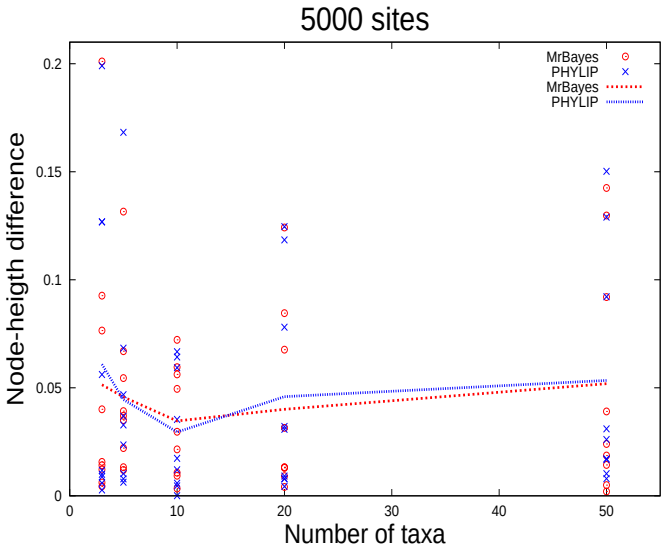

Figure 15: Node-height difference for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every number of taxa there are 9 observations for each of the programs.

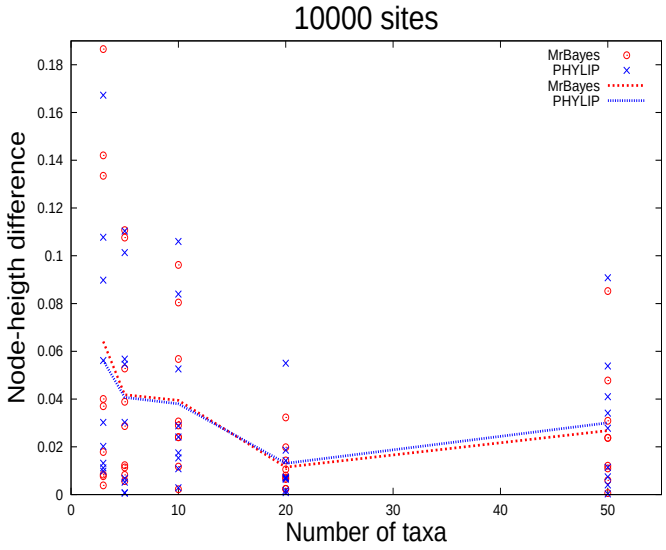

Figure 16: Node-height difference for MrBayes and PHYLIP where a dot represent a single observation and the line represent the mean values for the dots. For every number of taxa there are 9 observations for each of the programs.

7 Discussion

In this thesis we have simulated phylogenetic trees and DNA-sequences. The tree simulation model could be a good simplification of the real world. When simulating the trees, the birth and death rates were not randomly chosen and because we use Bayesian inference in MrBayes we should in a future study consider to randomize these rates from their prior distributions in MrBayes.

If we look at the model for DNA simulation we conclude that it is far from realistic. We have only considered one type of DNA-evolution, nucleotide substitution, but we know that the evolutionary process involves lots of more factors. The model for DNA-substitution, the Jukes-Cantor model, is also a rough simplification and we have for example not regarded the correlation between sites close to each other. But since we have used these simplified models in the analyses of both programs, we can compare the programs estimates and study how they depend on the number of taxa and the length of the sequences.

In our analyses, when we held the number of taxa constant and increased the sequence length, we can see that in most cases the estimates gets better with longer sequences. In figure 8, where we have three taxa, we did not see the same trend but in that case we probably have much information even with a sequence length of 1000 sites. It would have been good to analyse

sequences with less than 1000 sites for these trees. We would then probably see that the precision of the estimates increase when we increase the number of sites.

When we, for a fixed length of the sequence, increased the number of taxa we did not see the expected improvement of the estimates. Most of the trees with three taxa had their split near the tips. To investigate how the inference change in these cases we should have analysed some trees with the split closer to the root.

All the trees we simulated and analysed are normalized so that the root is at height 1. Although the time is on a relative scale we could measure the time in units of expected number of substitutions per site. The estimated branch length are in units of expected number of substitutions per site and by expressing our simulated tree in the same units we would not have to normalize the programs estimates.

As we can see in our analyses there where no significant difference between the estimates from both the programs. However, there was a tendency that PHYLIP performs better when increasing the sequences length with fixed amount of taxa. Another important difference is that the analyses in MrBayes took considerable more time than they did in PHYLIP. We would therefore recommend PHYLIP for analyses of datasets similar to the ones that we have analysed.

References

[1] Eaton, J. W. Gnu octave manual, 2002.

[2] Felsenstein, J. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol 17*, 6 (1981), 368–376.

[3] Felsenstein, J. Phylip source code and documentation, 1995.

[4] Harvey, P. H., May, R. M., and Nee, S. Phylogenies without fossils. *Evolution 48*, 3 (1994), 523–529.

[5] Hillis, D. M., Moritz, C., and Mable, B. K. *Molecular Systematics*. Sinauer Associates Inc, Canada, 1996.

[6] Huelsenbeck, J. P., and Ronquist, F. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics 17*, 8 (August 2001), 754–755.

[7] Moret, B. M. E., Roshan, U., Warnow, T., and Warnow, Y. Sequence-length requirements for phylogenetic methods, 2002.

[8] Rannala, B., and Yang, Z. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol 43* (1996), 304–311.

[9] Resnick, S. I. *Adventures in stochastic processes*. Birkhauser Verlag, Basel, Switzerland, 1992.

[10] Robinson, D. F., and Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences 53*, 1-2 (1981), 131 – 147.

[11] Svennblad, B., and Britton, T. Improving divergence time estimation in phylogenetics: More taxa vs. longer sequences. *Statistical Applications in Genetics and Molecular Biology 6*, 1 (2008), 35.

A Simulated trees

In Figure 17-19 we see the trees of size 100 that we used in our study. When the ratio $\frac{\mu}{\lambda}$ increases more species go extinct and we can see, by the enumeration, that there are 339 extinct species in Figure 19.

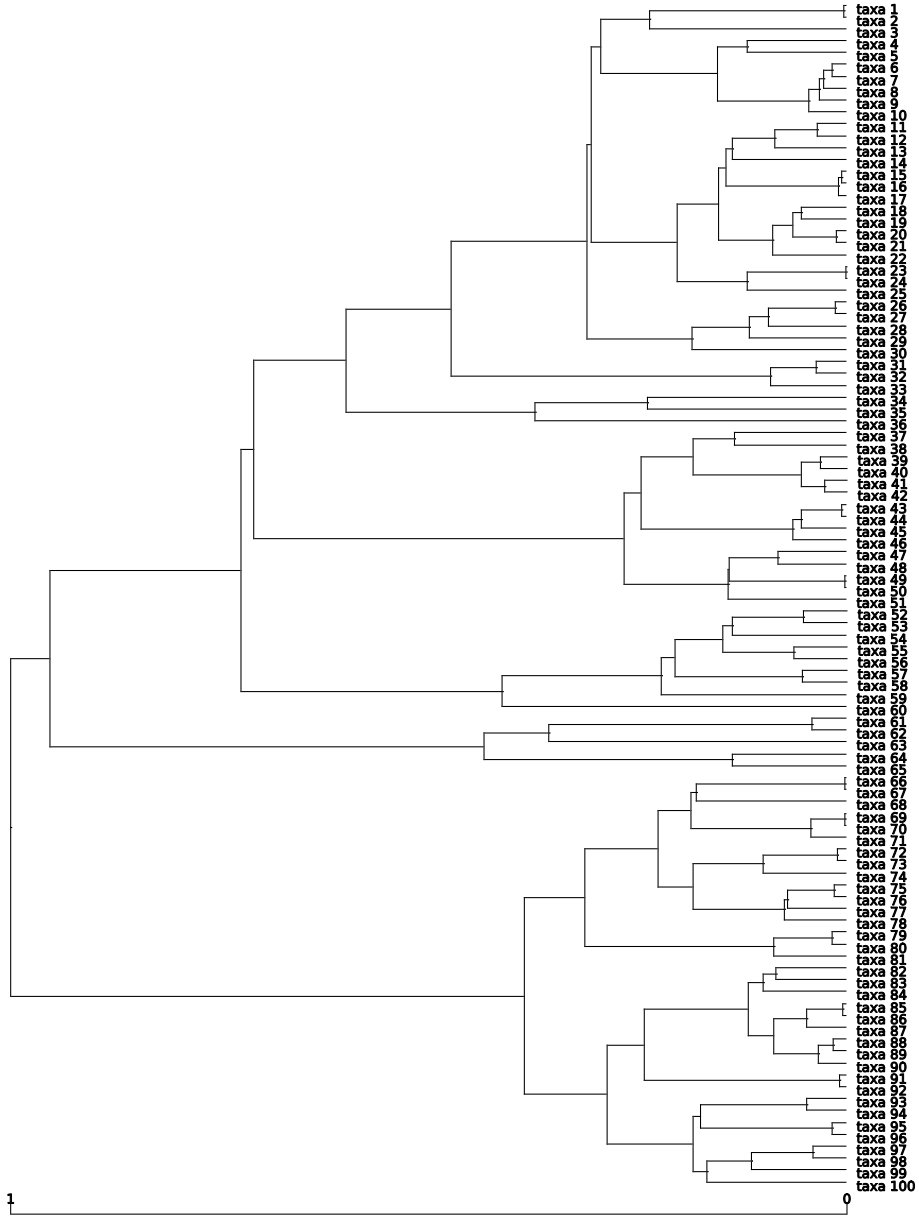

Figure 17: The simulated tree with $\frac{\mu}{\lambda} = 0$

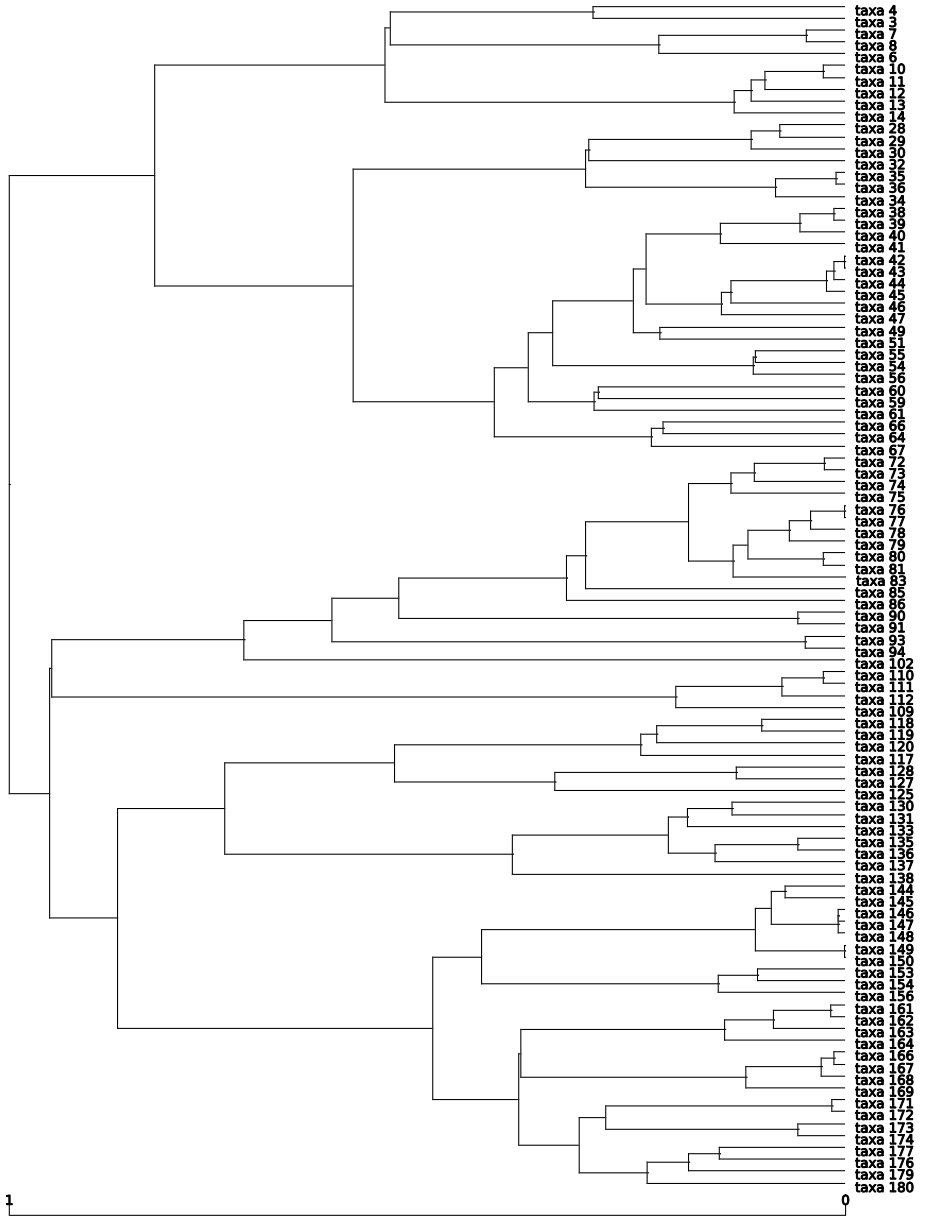

Figure 18: The simulated tree with $\frac{\mu}{\lambda} = 0.5$

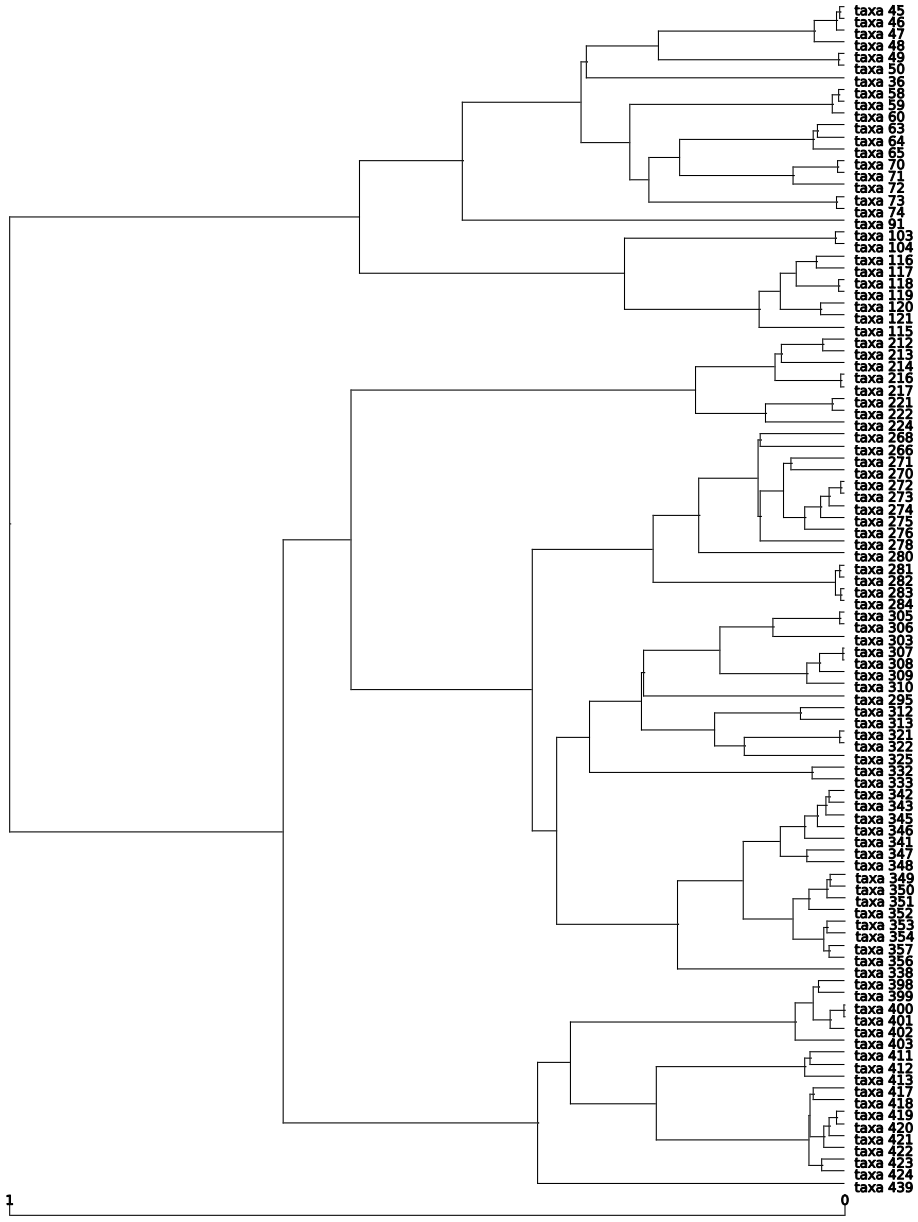

Figure 19: The simulated tree with $\frac{\mu}{\lambda} = 0.75$