RELIABILITY ANALYSIS OF RECONSTRUCTING PHYLOGENIES UNDER LONG BRANCH ATTRACTION CONDITIONS By

Ranjan Dissanayake, B.Sc.

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Statistics

University of Alaska Fairbanks

May 2018

APPROVED:

Elizabeth Allman, Committee Co-Chair
Julie McIntyre, Committee Co-Chair
Margaret Short, Committee Member
Scott Goddard, Committee Member
Leah Berman, Chair
Department of Mathematics and Statistics

Abstract

In this simulation study we examined the reliability of three phylogenetic reconstruction techniques in a long branch attraction (LBA) situation: Maximum Parsimony (MP), Neighbor Joining (NJ), and Maximum Likelihood. Data were simulated under five DNA substitution models–JC, K2P, F81, HKY, and GTR–from four different taxa. Two branch length parameters of four taxon trees ranging from 0.05 to 0.75 with an increment of 0.02 were used to simulate DNA data under each model. For each model we simulated DNA sequences with 100, 250, 500 and 1000 sites with 100 replicates. When we have enough data the maximum likelihood technique is the most reliable of the three methods examined in this study for reconstructing phylogenies under LBA conditions. We also find that MP is the most sensitive to LBA conditions and that Neighbor Joining performs well under LBA conditions compared to MP.

Acknowledgments

I would like to thank my advisor, Elizabeth Allman, for help and advice throughout the theoretical and computational modeling and writing process. I would also like to thank Shirley Li for helping me to correct my writing. Also I would like to recognize John Rhodes for teaching me phylogenetic theory and helping me to initiate this study. Thank you as well to my committee for help and constructive feedback.

Contents

List of Figures

List of Tables

1. Introduction

Charles Darwin, in 1859, described the origin of species using a tree diagram to represent the evolutionary relationship between ancestors of living species. Based on this concept, improved methods of studying evolutionary relationships using evolutionary trees, or phylogenies, have helped advance modern systematic biology and taxonomy. Phylogenetics is the study of the evolutionary history of species or the understanding of the ancestral relationships between species. A phylogenetic tree (phylogeny) is a tree diagram that contains nodes and edges without any cycles. Each node represents a species, and each edge represents the evolutionary process from ancestor to descendant through time. DNA sequences are often the basic data used to reconstruct the phylogenies.

DNA consists of the polynucleotide strands which have the form of a double helix. DNA includes nucleotides, or bases, which are adenine (A), guanine (G), cytosine (C), and thymine (T). Ordered arrangements of these bases are called DNA sequences and are denoted, for example, as $AAGTGGCTCC$. Based on the chemical similarities of DNA bases, adenine and guanine can be categorized as belonging to the purine group while cytosine and thymine can be categorized as belonging to the pyrimidine group. Moreover, A pairs with T while G pairs with C, owing to hydrogen bonds in the double helix structure of DNA. For example, if along one strand of the double helix there is a sequence of bases such as $AAGGGCTCC$, then the other strand contains the complementary sequence, $TTCCCGAGG$. Therefore, studying the existing information in DNA only requires sequences of one half of the helix. A piece of DNA molecule that determines a hereditary characteristic is known as a gene. Each individual has two copies of each gene which are inherited from their parents. Most of the genes are very similar among all humans. These small differences, or alleles, make a human being unique. A DNA mutation is a permanent change occurring in the DNA sequence of a gene. A common mutation that happens in DNA copying is base substitution. This happens when one base of the sequence replaces another base. A base substitution of purine to purine or pyrimidine to pyrimidine is called a transition while a substitution of purine to pyrimidine or pyrimidine to purine is called transversion. For example, suppose the ancestor's sequence $AAG\boldsymbol{T}G\boldsymbol{G}CTCC$ becomes $AAG\boldsymbol{C}G\boldsymbol{T}CTCC$ in a descendant. Then the base substitution that occurred at the fourth site, $T \rightarrow C$, is a transition, and the substitution that occurred at the sixth site, $G \rightarrow T$, is a transversion. DNA can also undergo other types of mutations such as insertions, deletions, and inversion of one or more consecutive bases. As these mutations are considered rare we assumed that the mutation process includes only base substitution. All these mutation processes evolving from ancestors to descendants are represented with trees.

Methods for reconstructing phylogenies have to consider mutation. Such methods are either character-based, distance-based or model-based. In character-based methods, all DNA sequences are simultaneously compared, and a score is calculated for each tree considering site-wise variation. A common technique used in this category is maximum parsimony (MP). In a distance method, evolutionary distances are calculated between all pairs of sequences and the resulting distance matrix is used for reconstructing a phylogenetic tree. In this category the most commonly used technique is the Neighbor Joining (NJ) method. Under the category of model-based methods Maximum Likelihood (ML) and Bayesian methods are used to estimate model parameters. The model-based method uses different probabilistic models to explain the evolutionary process under different conditions. Reconstruction methods MP, NJ, and ML and the probabilistic models, specifically JC, F81, K2P, HKY, and GTR we used in this study are discussed in detail in the background section. However, we can not always expect these methods to give us the true evolutionary tree. Long Branch Attraction (LBA) is one of the situations in which phylogenetic reconstruction techniques might have trouble inferring the "true" tree in phylogenetic inference, which is discussed below.

1.1. **Long Branch Attraction**

In this section we discuss the challenges that occur in reconstructing phylogenies due to long branch attraction. Figure 1 (a) illustrates the topology of a simple phylogenetic tree. Vertices which have only one adjacent edge are called leaves of the tree. Other vertices are internal vertices, and leaves represent taxa (species). Moreover, an edge incident to two interior vertices is called an interior edge, while one incident to a leaf is called a pendant edge. Usually the edge length in a phylogenetic tree is used to represent how many changes occur in sequences between the two ends of the edge, both seen and unseen. The phylogenetic tree which uses its edge lengths to explain the evolutionary process between nodes is called a metric tree. Therefore, a metric tree plays an important role in phylogenetic tree inference. An example of a metric tree is shown in Figure 1 (b) where the l_i are branch or edge lengths.

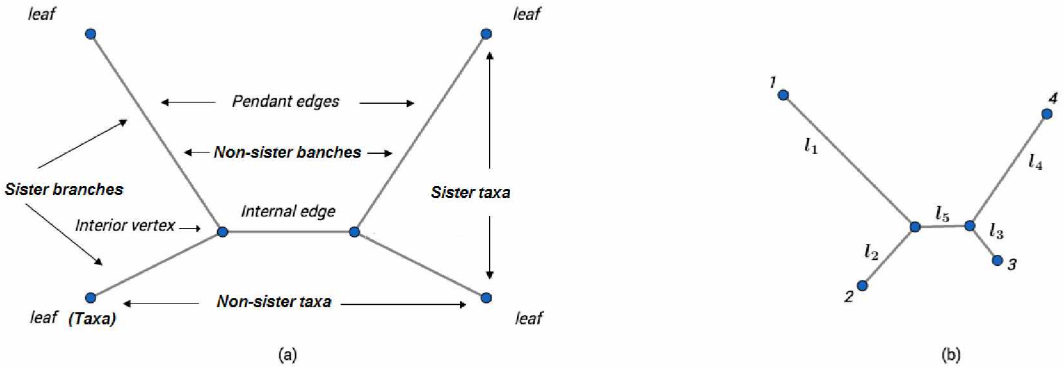


FIGURE 1. Four-taxon unrooted metric tree topologies. In 1 (b), taxa 1 and 2 are sister taxa and taxon 1 and taxon 4 are non-sister taxa. Further, it is clear that there are likely fewer sequence changes between taxon 3 and taxon 2.

Generally, phylogenetic trees are categorized into two groups: rooted and unrooted. A rooted tree is a directed tree with a vertex r called the root where all edges are directed away from it. An unrooted tree is a tree obtained after 1) suppressing the root from a rooted tree, and 2) ignoring all the directions. For example, Figure 2 shows a four-taxon unrooted tree which can be derived from the all five possible rooted trees. Their undirected edges can be used to illustrate the evolutionary clustering between taxa. All the phylogenies considered in this study are unrooted trees.

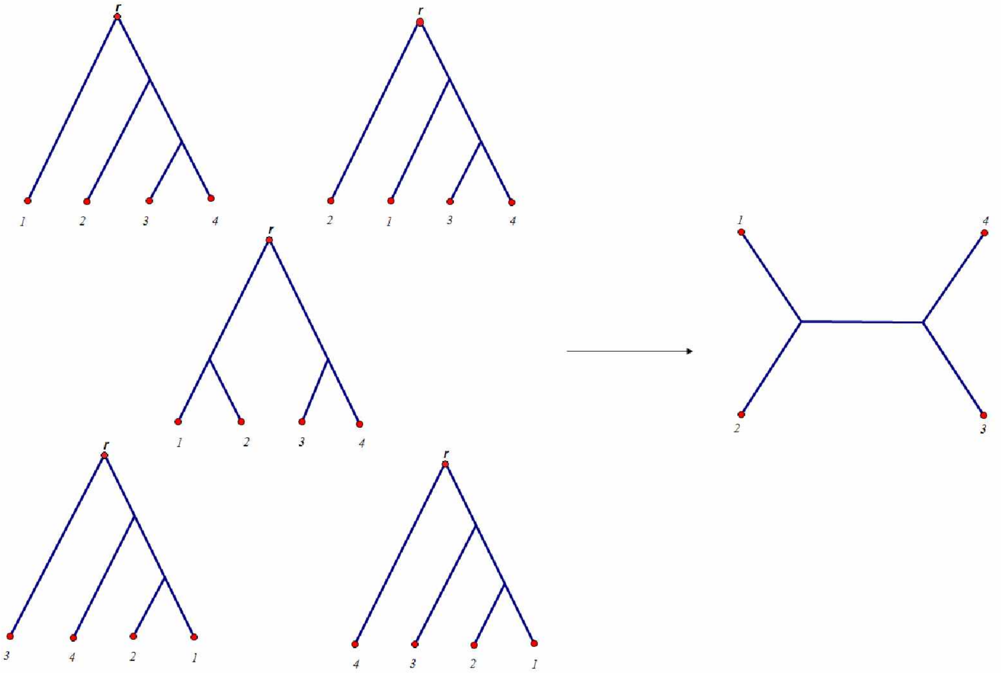


FIGURE 2. A four-taxon unrooted tree with all five possible rooted versions shown.

Figure 3 shows all the possible unrooted tree topologies for four taxa that reconstruction techniques can infer. It is clear that if we use one of the following trees to simulate DNA sequences, then there is a possibility for two wrong tree topologies to be inferred. Therefore the reliability of each reconstruction technique depends on the percentage of time that the correct tree is inferred.

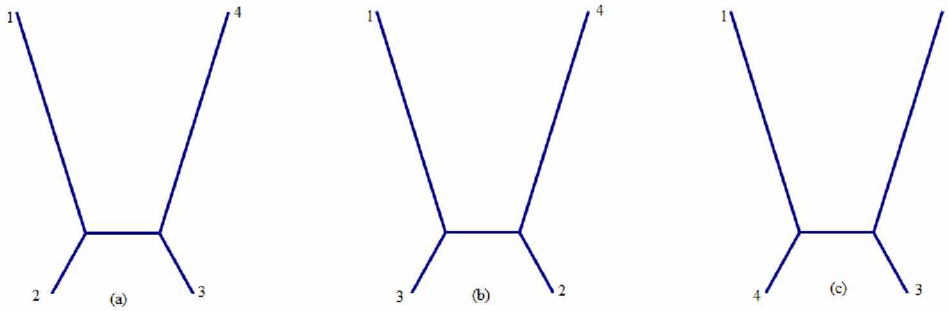


FIGURE 3. Four taxon unrooted tree topologies

An unrooted four-taxon metric tree with two non-sister long branches (B) and three other short branches (A) is shown in Figure 4 (a). When the difference between the lengths of two sets of branches increases, this metric topology makes it difficult to infer the correct tree because the two taxa that are most closely related metrically are not most closely related topologically. In other words, dissimilarities between the DNA sequences of sister taxa and similarities of the non-sister sequences add noise to the reconstruction process. This situation explains the long branch attraction (LBA) problem in reconstructing phylogenies. Therefore it is possible that the probability of inferring the wrong tree, as shown in Figure 4 (b), also increases.

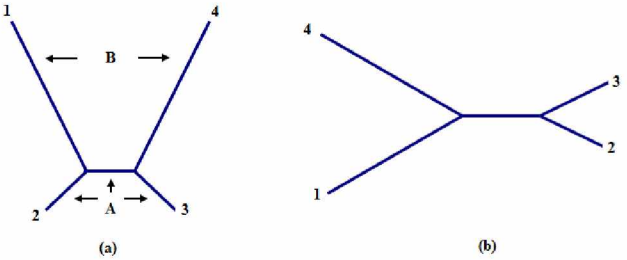


Figure 4. LBA metric tree topology where two non-sister taxa have long branches compared to the rest of the tree (a). A wrongly inferred tree with the long branches clustered is also shown (b).

1.2. **Objective**

The objective of this simulation study is to compare the reliability of the following phylogenetics techniques–MP, NJ, and ML–in estimating the correct tree under LBA conditions for DNA sequences simulated under different probabilistic models. Simulation studies are useful in phylogenetics because certain reconstruction techniques and models of the evolutionary process make assumptions. It is possible to simulate data under certain model assumptions and conditions that we are expecting to study. This simulation study can be used to examine the reliability of reconstruction techniques under LBA conditions.

This paper is organized as follows. In the background section, we present the phylogenetic theories and statistical models applied in this study. In the methods section, we explain how to simulate data and reconstruct phylogenies under different conditions. Next in the results and conclusion section we discuss how reconstruction techniques perform under long branch attraction conditions. Finally, in the discussion section we explain the future directions of this study.

2. Background

In this section we explain the background about probabilistic models and reconstruction techniques, which are used in this simulation study. First we discuss the Parsimony method, which is the simplest phylogenetic reconstruction technique used in phylogenetic analysis.

2.1. **Parsimony**

The major objective of this method is to select the best tree from data that minimizes the number of evolutionary changes or MP score. We consider the following aligned DNA sequences with the rows representing the sequences from the species. In general, we consider DNA bases in each column as sites (s_i); the data consists of five sites: s_1, s_2, s_3, s_4, and s_5. We observe that variations in characters are only at sites s_2 and s_3. These are the characters that explain the evolutionary relationship between these four taxa.

	s_1	s_2	s_3	s_4	s_5
1	A	T	T	C	G
2	A	T	G	C	G
3	A	G	C	C	G
4	A	G	G	C	G

We consider s_1, which has constant characters, to calculate the MP score. We observe that the *parsimony score* (the minimum number of mutations that can occur in a tree) for all possible trees for this site is zero. The site s_3 does have two or more different states. However, two of the states do not occur at least twice, which causes all the tree topologies to have the same MP score. Therefore those types of sites with such characteristics are considered as non-informative characters in parsimony analysis. In contrast to those, if we use s_2 to calculate MP scores of possible tree topologies, we observe that the parsimony score varies from tree to tree. Therefore if we consider any site which has at least two different states occurring at least twice; these are called informative characters in a parsimony analysis. We used informative characters to find the most parsimonious trees. Parsimony is usually performed on a rooted tree, but it does not matter whether the tree is rooted or unrooted. So in this case we have three different unrooted tree topologies for which we must compute parsimony scores. We apply the parsimony criterion to all three trees and find that one or more trees have the smallest parsimony score. Figure 5 shows how the MP score is calculated for s_2 on two different rooted trees. Finally, we select the tree or trees which results in the minimum parsimony score as the optimal tree.

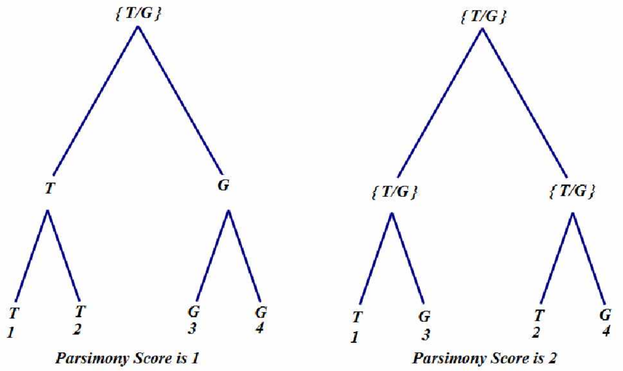


FIGURE 5. Computing the Parsimony score of a tree for one character constructs sets of states at internal nodes and yields a score of 1 for the tree (a) and 2 for the tree (b).

When n is large, this scenario is not easy to undertake and most software uses heuristic approaches to find an optimal tree or trees. When mutations are rare, MP is a justifiable method for inferring a tree. However, when mutations are not rare then probabilistic models and model-based methods are more commonly used.

2.2. **Probabilistic models in DNA mutation**

Let us consider the DNA sequence of an ancestor $S0$: $ATGCGCCATG$ and a descendant (species 1) sequence $S1$: $ATTCGCTGAA$. Based on the *iid* (identically and independently distributed) assumption, where each site in the DNA sequences behaves independently with an identical distribution, we can define the base distribution for any given base at $S0$ as $p_0 = (P(S0 = A), P(S0 = C), P(S0 = G), P(S0 = T)) = (p_A, p_C, p_G, p_T)$. For an example $p_0 \approx (\hat{p}_A, \hat{p}_C, \hat{p}_G, \hat{p}_T) = (2/10, 3/10, 3/10, 2/10)$ for the sequences above. Similarly, the base distribution at $S1$ can be written as $p_1 = (p_A, p_C, p_G, p_T)$. According to the probabilities of base substitutions, $S0$ evolves into $S1$, so we can observe that 16 possibilities such as $A \to A, A \to C, A \to G, A \to T, C \to A$ *etc.* could occur. These conditional probabilities can be represented with the following matrix. For instance, $A \to A$ means that an ancestral A remains an A in the descendant, and this probability is denoted by $P(S1 = A|S0 = A) = p_{AA}$. As an example, $\hat{p}_{AA}$ might be estimated to be 1/2 from the sequences above, because half of the A's in S_0 are still A's in S_1. The rows of the matrix refer to ancestral states, while the columns refer to descendant ones.

$$M = \begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AT} \\ p_{GA} & p_{GG} & p_{GC} & p_{GT} \\ p_{CA} & p_{CG} & p_{CC} & p_{CT} \\ p_{TA} & p_{TG} & p_{TC} & p_{TT} \end{pmatrix} \Longrightarrow \hat{M} = \begin{pmatrix} 1/2 & 0 & 1/2 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/2 & 0 & 0 & 1/2 \end{pmatrix}$$

Note that the product of p_0M gives the base distribution of $S1$, $p_0M = p_1$.

If the sequence $S1$ continues to evolve under the similar conditions of evolution from $S0$ to $S1$, and the elapsed time is similar, then we can write the distribution of base compositions at $S2$ as $p_2 = p_1M = p_0M^2$, where we are calling p_0 the root distribution. Assuming M represents the change after one time-step and models the sequence evolving through discrete time, we can model the sequence after n time steps by $p_n = p_0M^n$. In summary, the parameters of this discrete time model are a base distribution p_0 of non-negative numbers that sum to one, which gives three parameters (any three of p_A, p_C, p_G, p_T) and a Markov matrix M of non-negative numbers whose rows sum to one, giving twelve parameters (each row has three with i, j entry $P(S_1 = j|S_0 = i)$). This model M is called the *general Markov model* (GMM), in phylogenetic analysis.

Given a many-edged tree, we can find the *joint distribution* of states at the leaves of a tree. For example, let us consider the two edge tree, shown in Figure 6 (b), where $S1$ and $S2$ are two leaves of a root π. The M_{e_i} is the Markov matrix on the edge leading to Si where $i = 1, 2$. According to the possible patterns that can occur at sites, we can denote the joint distribution by a 4×4 matrix. Let us consider this situation: any base k at π must become an i in $S1$, and a j in $S2$. Therefore, the ij^{th} entry of the joint distribution

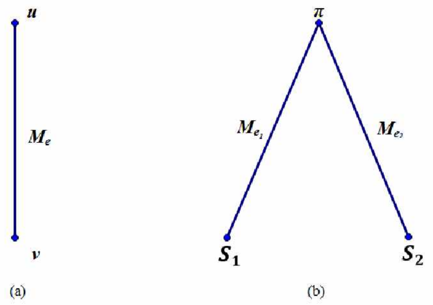


FIGURE 6. One edge tree (a) and two edge tree (b).

matrix can be written as,

$$\begin{aligned}P(i,j) &= \sum_{k=1}^{4} p_{\pi_k} M_{e_1}(k,i) M_{e_2}(k,j)\\ &= p_{\pi_1} M_{e_1}(1,i)M_{e_2}(1,j) + p_{\pi_2} M_{e_1}(2,i)M_{e_2}(2,j) + p_{\pi_3} M_{e_1}(3,i)M_{e_2}(3,j) + p_{\pi_4} M_{e_1}(4,i)M_{e_2}(4,j).\end{aligned}$$

Often, it is common to describe the evolutionary process using a continuous-time formulation. For this we introduce a rate matrix Q. Here q_{ij} denotes the instantaneous rate at which state i is replaced by j where $q_{ij} \geq 0, i \neq j$. Also, we require that the row sum of Q is 0:

$$Q = \begin{pmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{pmatrix}.$$

Let p_t denote the base distribution of four states at time t, where $t = 0$ gives the ancestral base distribution. Then we can write the following system of differential equations,

$$\frac{d}{dt} p_i(t) = p_i(t)q_{ii} + p_j(t)q_{ji} + p_k(t)q_{ki} + p_l(t)q_{li},$$

where i, j, k, and l are states and $i \neq j, k, l$.

Moreover we can denote those differential equations in matrix form as $\frac{d}{dt}\mathbf{p}_t = \mathbf{p}_t Q$. Then, solving the system of differential equations, using the given initial value p_0, we end up with the solution $p_t = p_0 e^{Qt}$. This result derives the relationship between the Markov matrix M and rate matrix Q as $M_e = M(t_e) = e^{Qt_e}$, where $M(t)$ is the single matrix describing the mutation along an edge representing a time of length t. Moreover, values of M are probabilities and values of Q are rates describing the instantaneous substitution process.

Tavare [1986] defined a model to be time-reversible if $\pi_i Q_{ij} = \pi_j Q_{ji}$ for any given time t. It follows that the probability of pattern ij is equal to the probability of pattern ji. The most general continuous-time model used in this simulation study is the *general time-reversible* (GTR) model introduced by Tavare [1986]. The GTR model on a tree is specified by the following parameters:

- an arbitrary choice of a root distribution $p = (p_A, p_C, p_G, p_T)$ where p is a probability mass function (pmf).
- an arbitrary choice of 6 rate parameters of $\alpha, \beta, \gamma, \delta, \epsilon, \eta$ with the common rate matrix Q on all edges,

$$Q = \begin{pmatrix} * & p_C\alpha & p_G\beta & p_T\gamma \\ p_A\alpha & * & p_G\delta & p_T\epsilon \\ p_A\beta & p_C\delta & * & p_T\eta \\ p_A\gamma & p_C\epsilon & p_G\eta & * \end{pmatrix}$$

  where the row sums are zero and $q_{AC} = q_{CA} = \alpha, q_{AG} = q_{GA} = \beta, q_{TA} = q_{AT} = \gamma, q_{CG} = q_{GC} = \delta, q_{GT} = q_{TG} = \epsilon, q_{GT} = q_{TG} = \eta$
- edge lengths of the tree

Five DNA substitution models have been used to explain the mutation process in this phylogenetic study and can be seen as restricted versions of the GTR model. JC is the most restrictive model and was formulated by Jukes and Cantor [1969]. In this model, we assume that all nucleotide substitutions rates are equal ($\alpha = \beta = \gamma = ... = \eta$) and all base frequencies are equal (uniform). The Kimura 2 parameter model (K2P), which was introduced by Kimura [1980], has the following setup. The base frequencies are assumed to be uniform and two nucleotide substitution types are allowed: transitions and transversions. The F81 model is the same as JC in that all nucleotide substitutions are equal, but has arbitrary base frequencies; this model was formulated by Joseph Felsenstein [1981]. Subsequently, the HKY model was formulated by Hasegawa et al. [1985] as an extension of the K2P model. In this model the base frequencies are non-uniform like in the F81 model. The most commonly used continuous time model used in data analysis is the GTR.

2.3. **Model-based Distances**

Let us consider two DNA sequences. If both sequences are identical, we say there is no dissimilarity between the two sequences or that there is 0 distance between the two sequences. Dissimilarity can be calculated from data by simply calculating the proportion of sites that are dissimilar in the two sequences. This distance is called “Hamming distance”. However, phylogenetic reconstruction techniques use different dissimilarity maps using a probabilistic model of molecular evolution, rather than the Hamming distances. In this simulation study, we use Jukes-Cantor Distance to reconstruct the phylogenies. Next, we consider the ancestral sequence $S0$, which has a uniform base distribution. Its mutation is explained by a Jukes-Cantor rate matrix:

$$Q = \begin{pmatrix} -\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & -\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & -\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & -\alpha \end{pmatrix}$$ where α is the rate at which any given base is replaced by a different base.

Then the total mutation process over the elapsed time t can be described by:

$$M_{(t)} = e^{Qt} = \begin{pmatrix} 1-a & a/3 & a/3 & a/3 \\ a/3 & 1-a & a/3 & a/3 \\ a/3 & a/3 & 1-a & a/3 \\ a/3 & a/3 & a/3 & 1-a \end{pmatrix} \text{ where } a = a(t) = \frac{3}{4}(1 - e^{-\frac{4}{3}\alpha t}).$$

We can find the joint distribution of ancestor descendant states after time t by:

$$P = \mathrm{diag}(p_0)M(t) = \begin{pmatrix} (1-a)/4 & a/12 & a/12 & a/12 \\ a/12 & (1-a)/4 & a/12 & a/12 \\ a/12 & a/12 & (1-a)/4 & a/12 \\ a/12 & a/12 & a/12 & (1-a)/4 \end{pmatrix}$$ where $P(i,j)$ gives the probability of an ancestral i and descendant j appearing at a site.

The off-diagonal entries are the frequencies of the various ways the states may disagree at a site in the two sequences. We can estimate a using the Hamming distance as $\hat{a} = \dfrac{\textit{number of sites that show different states}}{\textit{total number of sites}}$. Finally we can estimate the total amount of mutation by $\widehat{\alpha t} = -\dfrac{3}{4}\ln(1 - \frac{4}{3}\hat{a})$ which is called the Jukes-Cantor distance $d_{JC}(S0, S1)$ between sequences $S0$ and $S1$. Likewise, we can calculate the model-based distances for other models too.

2.4. Neighbor Joining

The Neighbor Joining Method is the most widely used distance-based method for phylogenetic reconstruction. The NJ method reconstructs a metric phylogeny. In each iteration of the algorithm, two nodes of the tree are chosen and defined as neighbors in the tree. The joining is done recursively until all of the nodes are paired together. In this simulation study we use JC distances to reconstruct NJ trees for each simulation.

First we calculate the dissimilarities between N taxa by JC distance. Calculated distances are stored in a symmetric $N \times N$ matrix or table D. If S_i, S_j taxa are to be joined in the current iteration then, we introduce a new vertex v to make them a sister taxa in a metric tree. Next we put the rest of the $N-2$ taxa into one temporary group S_g and find the branch lengths from S_i and S_j to vertex v by solving the following system of equations (Figure 7 will help you to understand the concept discussed here). Then we replace S_i, S_j with v in the distance table and iterate. We proceed with this method until all taxa have been joined in a tree, the NJ tree.

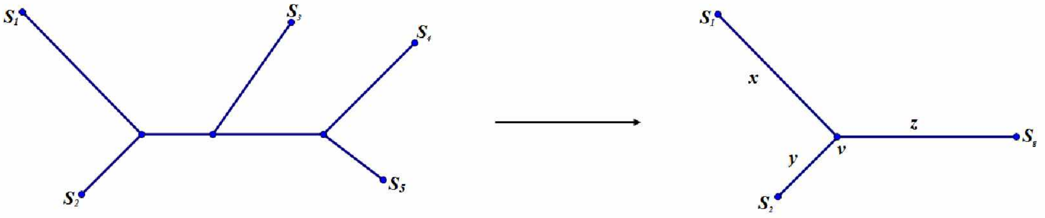


FIGURE 7. Distance between S_1 and S_2 denoted by $d_{1,2}$, which is the value of element $N_{1,2}$ in the D matrix. The right hand side graph shows the situation we used to calculate the branch lengths by $x + y = d_{1,2}$, $x + z = d_{1,3}$, $y + z = d_{2,3}$. Here x and y are the true branch lengths presented in the final NJ tree.

2.5. **Maximum Likelihood (ML)**

The probabilistic models we use to explain evolutionary process have parameters. The main idea behind phylogeny inference with maximum likelihood is to determine the tree topology, branch lengths and numerical parameters of the evolutionary model by maximizing the likelihood function. Let us consider a Jukes-Cantor model on a one-edge tree, from an ancestral sequence $S0$ to a descendant sequence $S1$. Let the length of the edge be t, measured in units so that $\alpha = 1$ in the Jukes-Cantor rate matrix discussed above. Then a becomes $a = \dfrac{3}{4}(1 - e^{-\frac{4}{3}t})$ and we need to estimate parameter t for the sequence data of $S0$ and $S1$. Next we summarize the sequence data by counting how often each pattern appears, such as $N(i, j) = n_{ij}$ where n_{ij} is the number of sites with base i in $S0$ and base j in $S1$. This N is a 4×4 matrix of data counts. Then using the joint distribution $P = (p_{ij})$, we can write the likelihood function as

$$L(t|n_{ij}) = \prod_{i,j=1}^{4} p_{ij}^{n_{ij}}$$

where p_{ij} is a function of parameter t (through function $a(t)$) and the log-likelihood is

$$\ln L(t|\{n_{ij}\}) = \sum_{i,j=1}^{4} n_{ij} \ln\ p_{ij} = \ln(a/12) \sum_{i \neq j} n_{ij} + \ln((1-a)/4) \sum_{i} n_{ii}.$$

Then, differentiating this with respect to t and setting this equal to 0, we can find the ML estimator of

$$a(\hat{t}) = \frac{\sum_{i \neq j} n_{ij}}{\sum_{ij} n_{ij}}.$$

This implies that ML's estimated edge length t is actually the JC distance. However, estimating a best maximum likelihood tree or trees is computationally expensive for a large number of taxa and sequences available because the ML procedure must consider all the topologies and find the best edge lengths for each topology.

3. Methods

The general approach of this simulation study is to explain how LBA conditions affect the reliability of three phylogenetic reconstruction techniques: MP, NJ, and ML on four-taxon trees. First we choose a model tree and parameters to simulate DNA sequences under five probabilistic phylogenetic models. Next we use three different phylogenetic techniques to reconstruct the tree. Finally, we calculate the percentage of the time that the correct tree is inferred from tree construction for each method.

3.1. **Model Tree**

The four-taxon unrooted tree topology shown in Figure 3 (a) is the model tree for this study. In each iteration of this simulation study we consider topologically the same model tree and vary the two branch length parameters A and B. Branch length parameters A and B range from 0.005 to 0.75 with an increment of 0.02 to simulate DNA data on these model trees. Figure 8 shows how model tree topologies with varying branch length combinations will appear in the branch length parameter space.

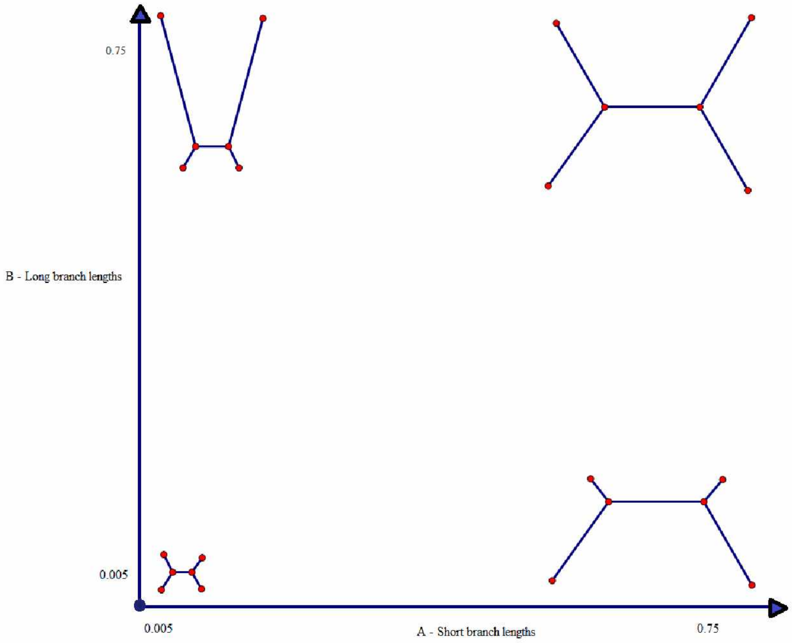


Figure 8. In branch length space, short branch lengths (A) are plotted on the horizontal axis and long branch lengths (B) are plotted on the vertical axis.
Source : Hulsenbeck, J.P. “Performance of Phylogenetics Method in Simulation”. *Systematic Biology, vol. 44, No.1, Mar.,1995*, pp.17-48.

3.2. **Simulate DNA Sequences**

The five different models of DNA substitution discussed in the background section were used to simulate DNA sequences in this study. For each model we simulated DNA sequences with 100, 250, 500 and 1000 sites with 100 replicates in each case. Table 1 explains models and parameter values used to simulate the sequences. We used `Seq-Gen Version 1.3.3` updated by Rambaut and Nick [2011] to generate the DNA sequences in this study.

Model	Base Frequencies	Substitution Rates	Number of Parameters
JC	$p_A = p_C = p_G = p_T$ $= 0.25$	$q_{AC} = q_{AG} = q_{AT} = q_{CG} =$ $q_{CT} = q_{GT} = \frac{1}{3}$	5
K2P	$p_A = p_C = p_G = p_T$ $= 0.25$	$q_{AG} = q_{CT} = 2,$ $q_{AC} = q_{AT} = q_{CG} = q_{GT} = 1$	$5 + 1 = 6$
F81	$p_A = 0.35, p_C = 0.15,$ $p_G = 0.25, p_T = 0.25$	$q_{AC} = q_{AG} = q_{AT} =$ $q_{CG} = q_{CT} = q_{GT} = \frac{1}{3}$	$5 + 3 = 8$
HKY	$p_A = 0.35, p_C = 0.15,$ $p_G = 0.25, p_T = 0.25$	$q_{AG} = q_{CT} = 2,$ $q_{AC} = q_{AT} = q_{CG} = q_{GT} = 1$	$5 + 4 = 9$
GTR	$p_A = 0.35, p_C = 0.15,$ $p_G = 0.25, p_T = 0.25$	$q_{AC} = 2, q_{AG} = 4, q_{AT} = 1.8,$ $q_{CG} = 1.4, q_{CT} = 6, q_{GT} = 1$	$5 + 8 = 13$

TABLE 1. Five models of nucleotide substitution and parameter values used to simulate DNA sequences. The same number of branch lengths (5) needs to be estimated for every model. Here p_A, p_G, p_C, p_T are the relevant probabilities in the pmf of the root distribution and $q_{i,j}$ are the substitution rates of state i to j in the rate matrix Q.

3.3. **Methods of Reconstructing Phylogeny**

In this study we employed three commonly used methods for reconstructing phylogenies. All these phylogenies were computed using `PHYLIP` programs which were originally developed by Joseph Felsenstein [1984]. First we used the parsimony technique available in the `dnapars` program to reconstruct the original tree. Next we used the `dnadist` program to calculate JC distances and used the `neighbor` program to reconstruct the original tree from the Neighbor Joining method. Finally we used the JC model setup in `dnaml` program to reconstruct a Maximum Likelihood tree.

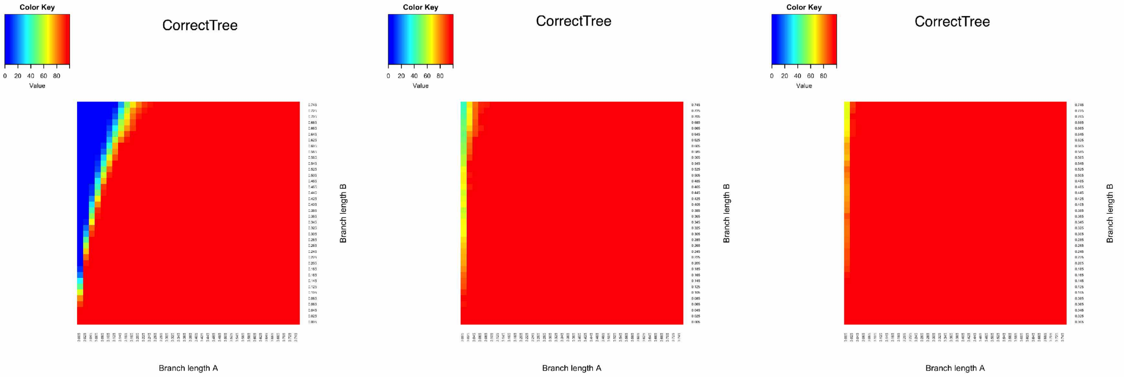
3.4. **Method of Evaluation**

We calculated the percentage of inferred tree topologies that are the same as the model tree in each case. According to the model tree topology used in Figure 3 (a), we categorized the tree topologies of Figure 3 (b) and (c) as wrong tree topologies. Among these, Figure 3 (c) shows the wrong tree topology that matches the topology expected when long branch attraction is present. Here we labeled trees in Figure 3 from left to right as “(a) - Correct Tree”, “(b) -Wrong Tree”, “(c) - LBA Wrong Tree”. In this study we replicated 100 data sets in each case. Then we calculated the percentage of inferred trees that were in each of these three categories. We calculated the Robinson Foulds (RF) distance between the reconstructed tree and the model tree for the tallying. A RF distance equal to 0 implies that the compared topologies are identical and that the reconstructed tree was counted in the relevant category. Finally we used heat maps to present the results. The results of the analysis are plotted as heat maps using the same axes as shown in Figure 5. The RF distance calculations and heat map constructions were done using `RStudio Version1.0.136`.

4. Results and Conclusion

The detailed results of this study are included in the Appendices. The figures show the reconstruction performance of parsimony, Neighbor Joining, and Maximum likelihood methods for 100, 250, 500, and 1000 sites. Appendices 1, 2, 3, 4, and 5 show the simulation results for the sequences generated under models JC, F81, K2, HKY, and GTR, respectively. Areas of the branch length parameter space colored red (hot color) estimate that 100% of the inferred trees are correct. The areas colored blue (cool color) indicate incorrectly estimated correct trees. Intermediate performances of reconstruction from correct to incorrect are indicated from hot colors to cooler colors (or red to blue).

First, we consider the performance of the three reconstruction techniques for sequences simulated under the JC model. The length of the generated sequences is 1000 sites. In Figure 9 (a), the top left portion of the branch length space shown in blue indicates a wrongly estimated correct tree. For tree topologies in this region, branch lengths A are very small and branch lengths B are very long which is the LBA region of branch length space. This implies that the MP method has difficulty inferring the correct tree topology in this region. By contrast, NJ and ML perform well in this region. Intermediate cooler colors show that NJ and ML methods also struggle to infer correct tree topologies in the top left corner, although the regions for NJ and ML are small. Comparing NJ and ML is not easy; by comparing the very top left corners of the Figure 9 (b) and 9 (c), the hot colors shows that ML performs better than NJ.

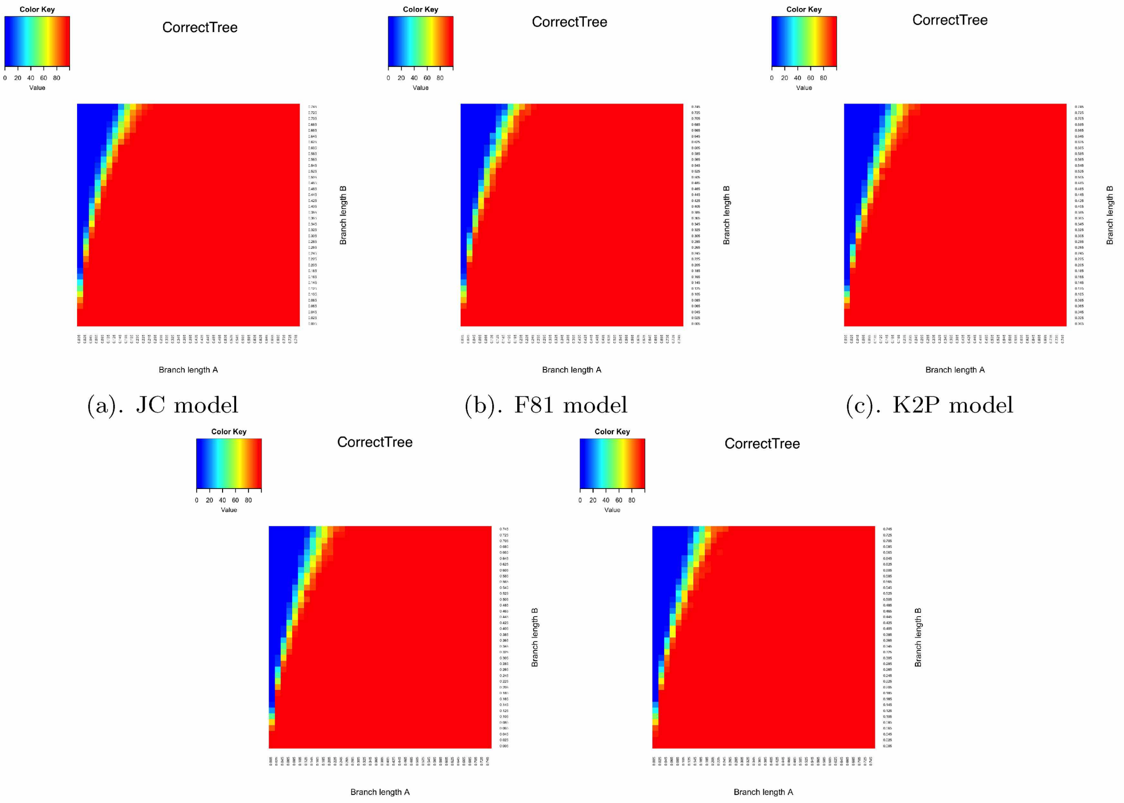


(a). Parsimony (b). NJ-(JC distance) (c). ML - (JC model)

Figure 9. Sequence length is 1000 (Sequences generated under JC model). Here the NJ method used JC distances and ML used JC model parameters to reconstruct the phylogenies.

Next we measure how the performance of each technique varies when decreasing the sequence length for data generated under the JC model. The first column of Appendix 1 shows how the performance of the parsimony technique varies when the sequence length changes from 100 to 1000. The blue color region which appears in the top left portion of the heat maps implies that for all sequence lengths, MP struggles to infer the correct tree topology in the LBA region. For sequence lengths of 500 and 1000, both NJ and ML perform well compared to MP. When the sequence length is too short (less than 500), NJ and ML also have at least some difficulty in inferring the correct tree at the extremes.

Figure 10 clearly shows that the MP method performs unreliably in the top left corner for each model. Therefore we can conclude that the MP method performs poorly in the LBA region of the edge length parameter space. According to Appendices 1-5, we can observe that NJ struggles to infer the correct tree topology at short sequence lengths for all these models. In the JC and F81 models, we noticed that when sequence lengths were 500 and 1000, NJ estimates the top left region more reliably. In the rest of the models, NJ clearly shows somewhat poor performance in this region, but performance is better when sequence lengths are large. Compared to the MP method, the region of poor behaviour for NJ is small. The ML method also has trouble correctly inferring tree topologies in this region when sequence lengths are very small. This is because when sequence length is small, less data is available for the ML method. Finally, when sequence lengths are large enough, ML performs better and more reliably in this region compared to the other two methods.

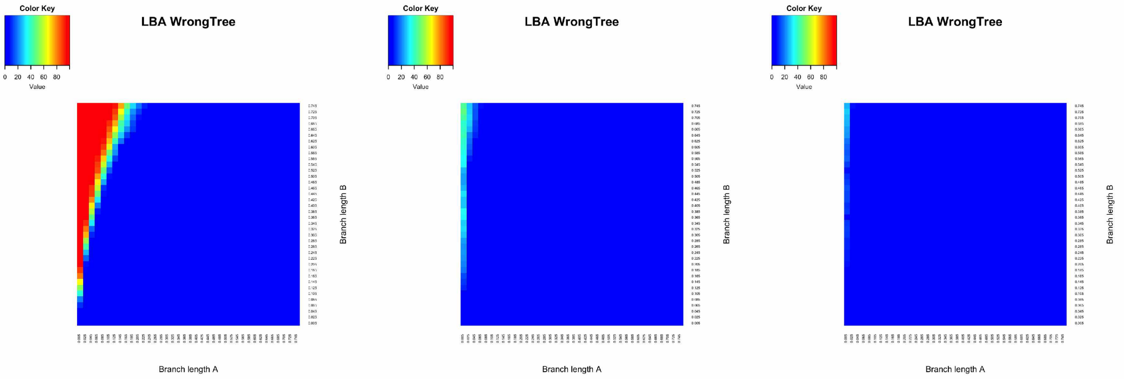


(a). JC model (b). F81 model (c). K2P model

(c). HKY model (c). GTR model

Figure 10. Sequence length is 1000 (Sequences generated under different models and reconstructed under MP method)

Figure 11 shows that all the failures explained by the correct tree inferred in Figure 9 came from the LBA wrong trees. For example, in Figure 9 (a) we see blue in the LBA region of branch length space indicating the true tree is not inferred, but instead MP selects the LBA wrong tree (see Figure 11 (a)) with red. Therefore we can clearly observe that the difficulty in inferring the correct tree happens for MP due to the LBA nature of the metric tree topology. Further it shows that the reliabilities of NJ and ML are also somewhat affected by the LBA condition. The effect of the LBA condition on NJ is small, while the effect on ML is negligible.



(a). Parsimony (b). NJ (c). ML

Figure 11. Sequence length is 1000 (Sequences generated under JC model)

Next we examine the reliability shown by NJ in Appendix 7. Here we can observe that the effect of the LBA conditions is smaller when the sequence length is 500 and 1000. Interestingly, when the sequence length is too small (< 500), there is a considerable effect on reliability at the top right corner of the branch length space which is not explained by LBA conditions. Finally, the ML results in Appendix 8 show that the reliability of the ML method is strong and the effect of the LBA is negligible. When sequence length is really small (< 500), ML shows some difficulties in inferring the correct tree but this is not entirely caused by LBA.

According to the above results the parsimony method is the most sensitive to LBA conditions, and Neighbor Joining performs well under LBA conditions compared to parsimony. Finally Maximum Likelihood is the best technique among the three methods we examined in this study. ML performs well under the LBA conditions when sequence lengths are long enough (greater than 500). ML handles Long Branch Attraction conditions really well compared to NJ and MP.

5. Discussion

In this study we used only JC models and distances for reconstructing trees. One of the issues that should be considered in the future is if we can extend this study by reconstructing trees under different models and distances. For example, when reconstructing phylogenies for DNA sequences simulated under the K2P model we should use K2P distance and the K2P model to reconstruct trees. Another concern that we can address is changing rate parameters and increasing sequence lengths to examine the performance of the reconstruction techniques. Also for further work we could use Bayesian methods to estimate the model parameters and reconstruct the trees under LBA conditions and compare that with Maximum likelihood results.

6. References

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

Felsenstein J., 1984. `PHYLIP Version 3.695` *http://evolution.genetics.washington.edu/phylip.html.*

Hasegawa, Kishino K., Yano T., 1985. “Dating the human-ape splitting by a molecular clock of mitochondrial DNA”, *Journal of Molecular Evolution*, vol. 22 (pp. 160-174).

Helsenbeck J.P., 1995. “Performance of phylogenetics method in simulation”, *Systematic Biology*, vol. 44 (pp. 17-48).

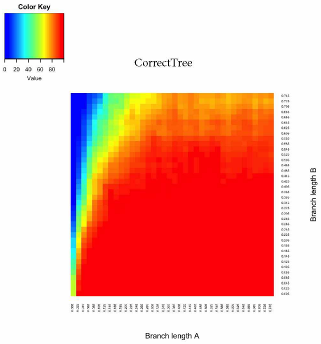
Jukes T.H., Cantor C.R., 1969. “Evolution of protein molecules”, *Mammalian protein metabolism III* (Munro H.M., ed.), New York Academic Press, (pp. 21-132).

Kimura M., 1985. “A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences”, *Journal of Molecular Evolution*, vol. 16 (pp. 111-120).

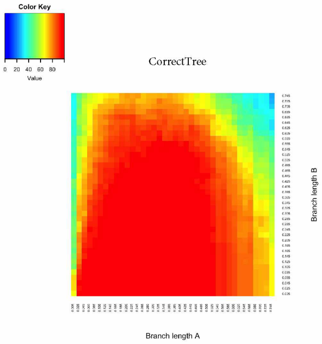
Rambaut A., Grassly N., 2011. `Seq-Gen Version 1.3.3` *http://tree.bio.ed.ac.uk/software/seqgen*

Tavare S., 1986. “Some probabilistic and statistical problems in the analysis of DNA sequences”. *Some mathematical question in biology - DNA sequence analysis* , Lectures on Mathematics in the Life Sciences, American Mathematical Society, (pp. 57-86).

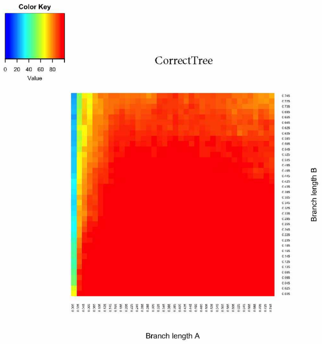
Appendix 1 : Sequences generated under JC model



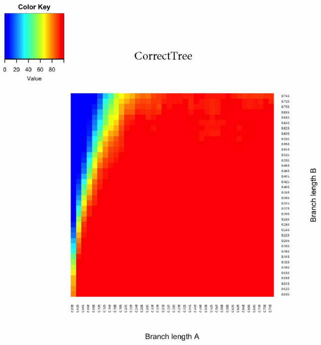
Parsimony-100



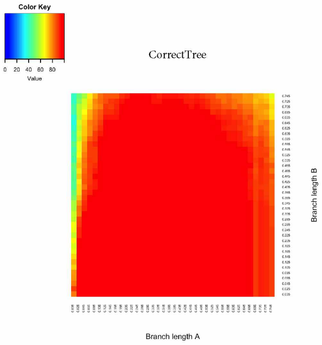
NJ-100



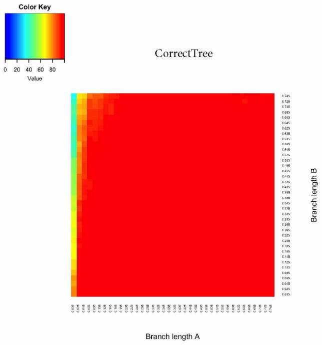
ML-100



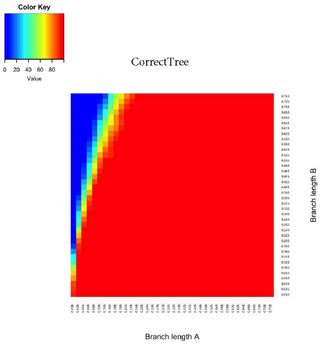
Parsimony-250



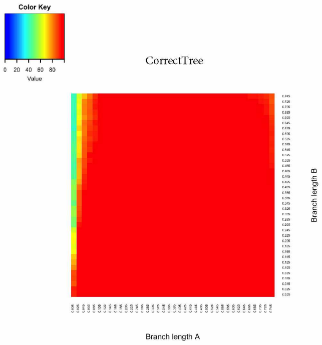
NJ-250



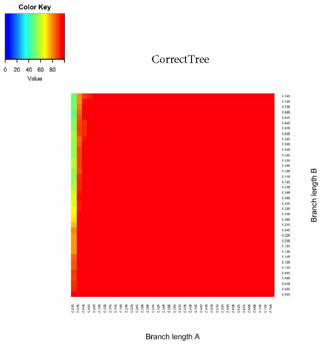
ML-250



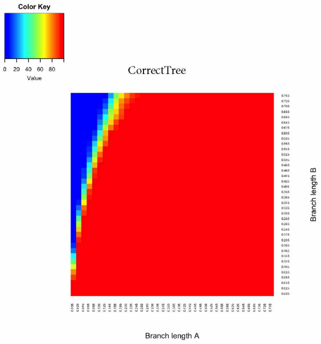
Parsimony-500



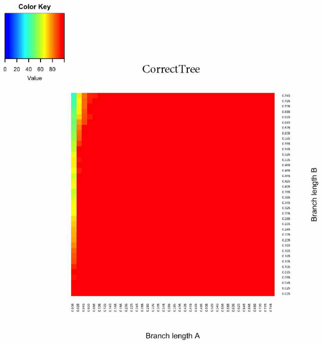
NJ-500



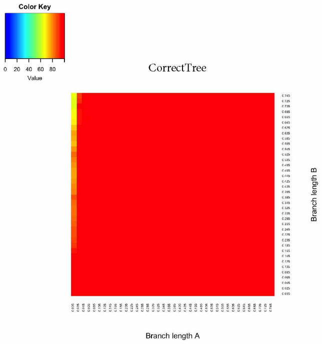
ML-500



Parsimony-1000

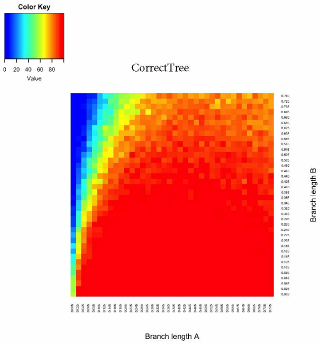


NJ-1000

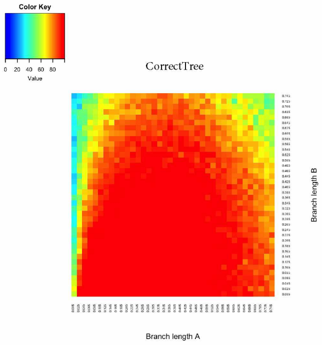


ML-1000

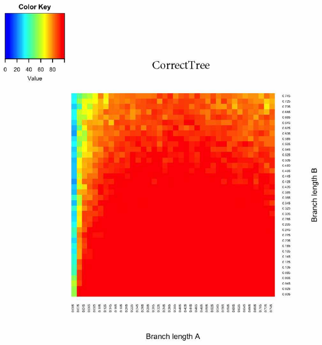
Appendix 2 : Sequences generated under F81 model



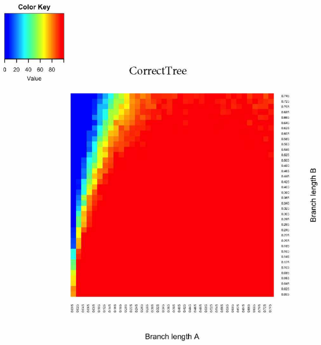
Parsimony-100



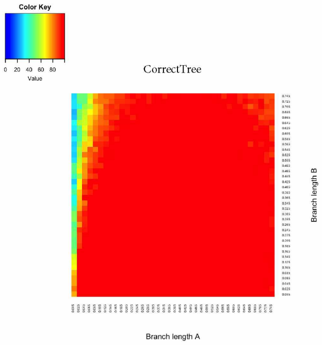
NJ-100



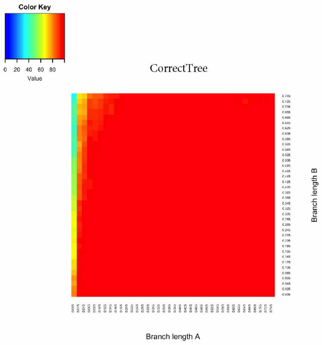
ML-100



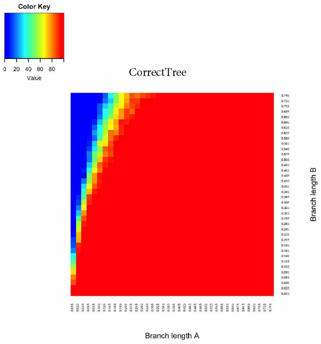
Parsimony-250



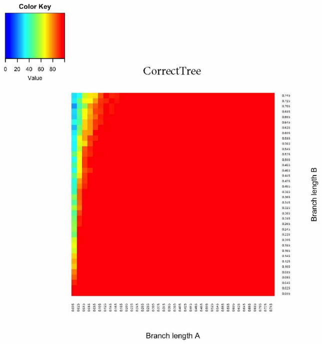
NJ-250



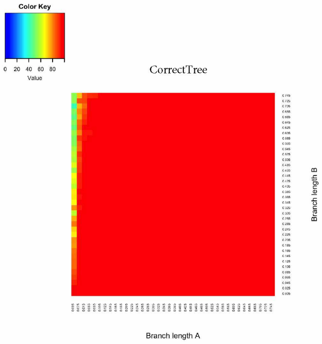
ML-250



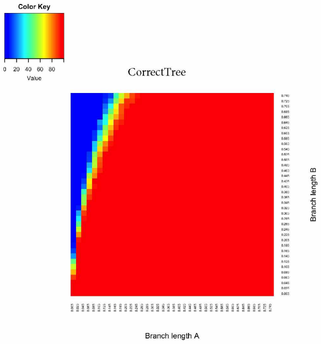
Parsimony-500



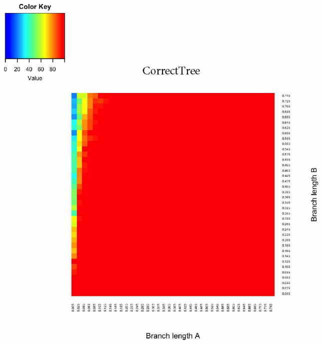
NJ-500



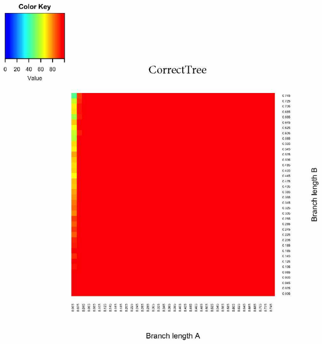
ML-500



Parsimony-1000

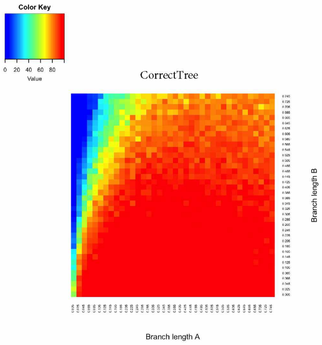


NJ-1000

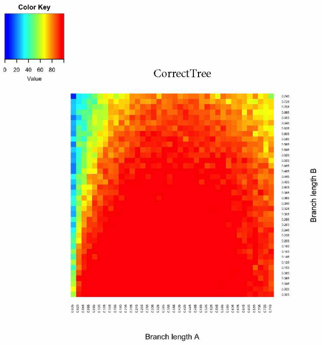


ML-1000

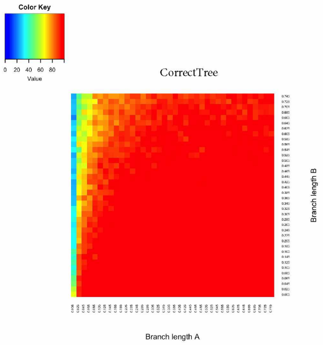
Appendix 3 : Sequences generated under K2P model



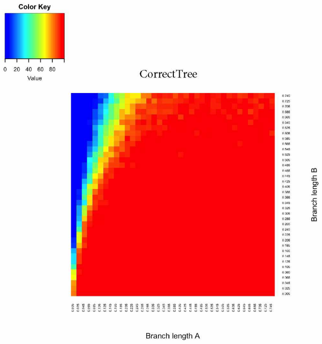
Parsimony-100



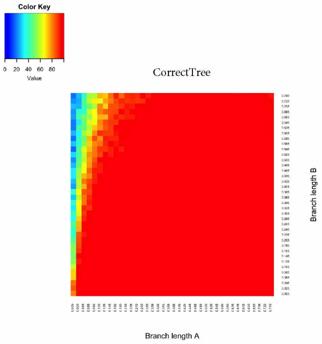
NJ-100



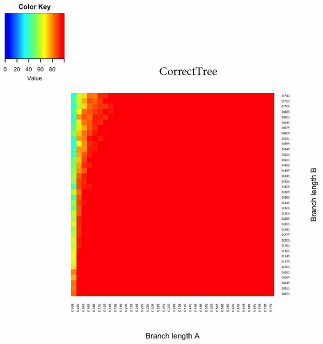
ML-100



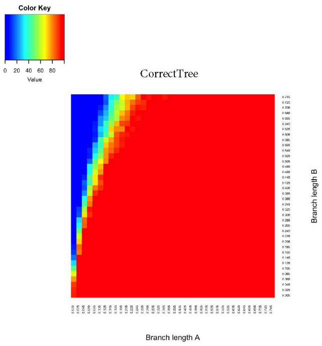
Parsimony-250



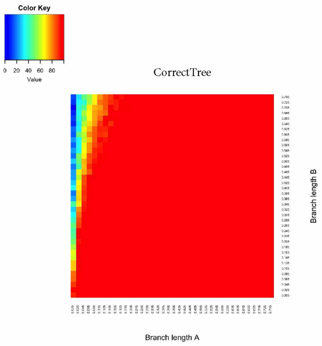
NJ-250



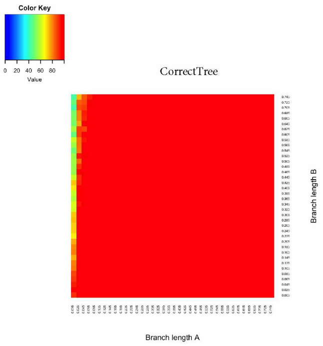
ML-250



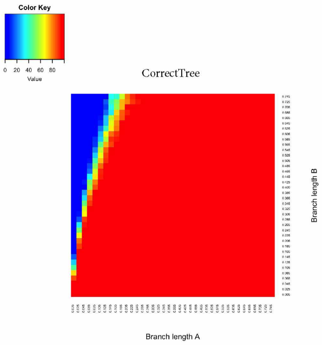
Parsimony-500



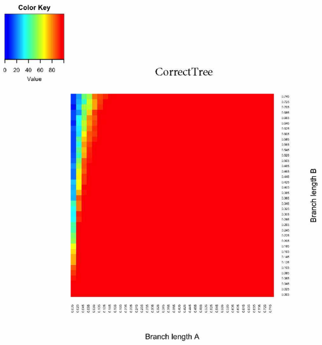
NJ-500



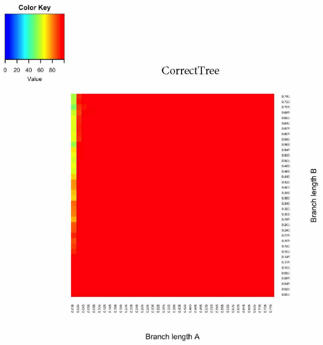
ML-500



Parsimony-1000

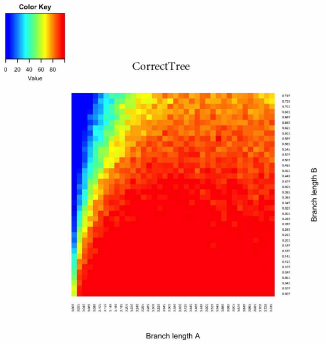


NJ-1000

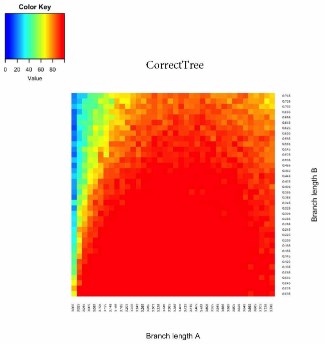


ML-1000

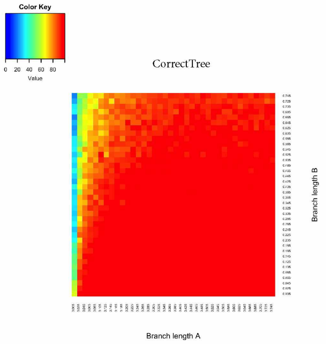
Appendix 4 : Sequences generated under HKY model



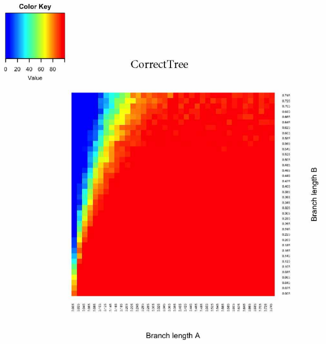
Parsimony-100



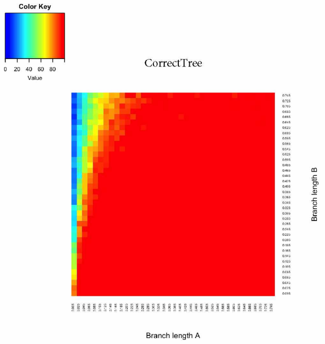
NJ-100



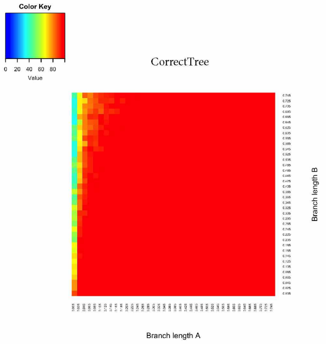
ML-100



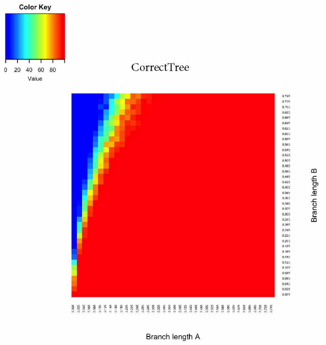
Parsimony-250



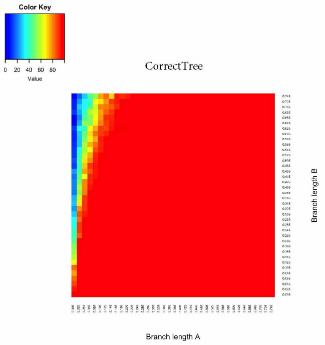
NJ-250



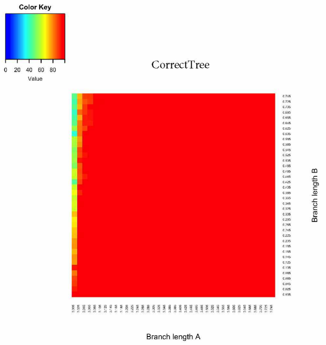
ML-250



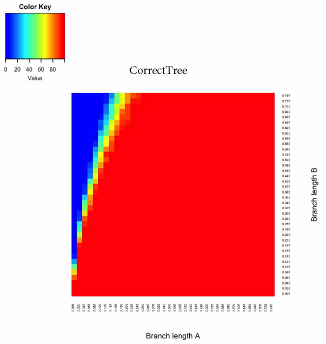
Parsimony-500



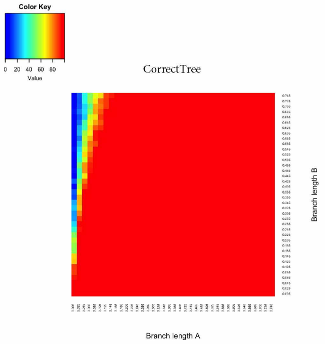
NJ-500



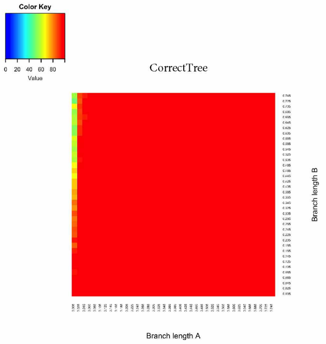
ML-500



Parsimony-1000

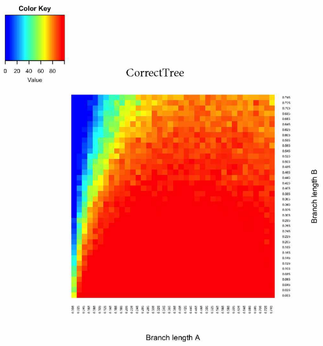


NJ-1000

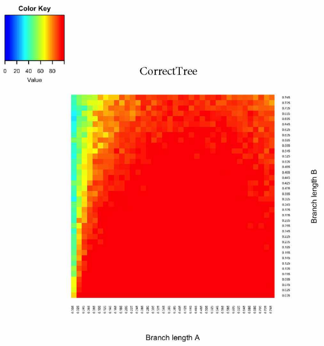


ML-1000

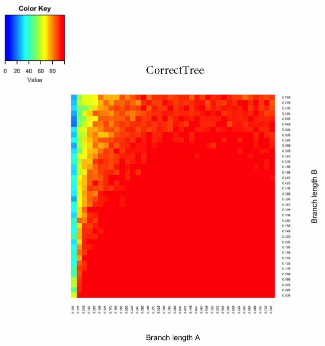
Appendix 5 : Sequences generated under GTR model



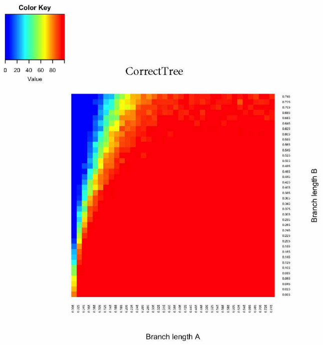
Parsimony-100



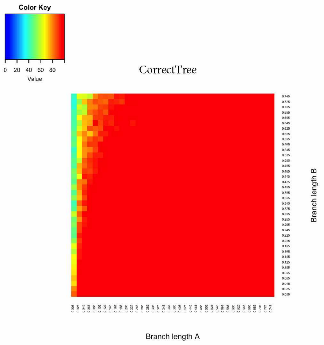
NJ-100



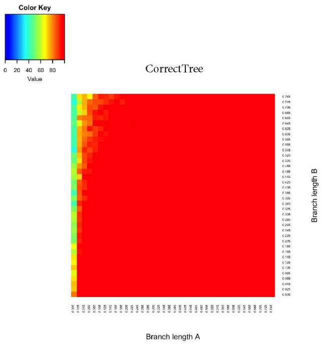
ML-100



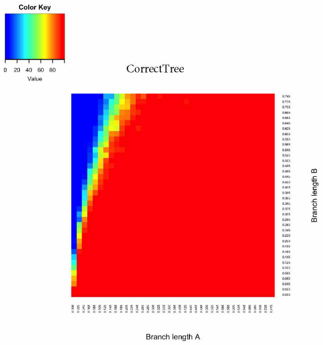
Parsimony-250



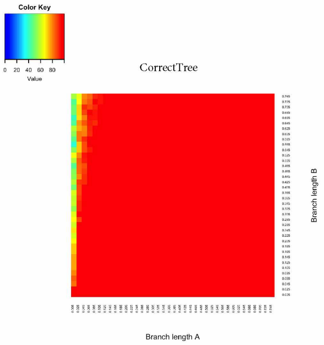
NJ-250



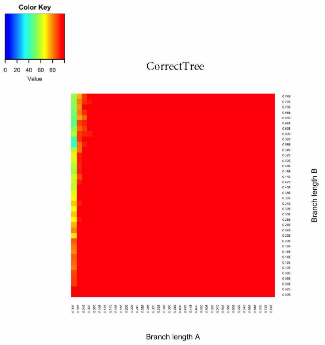
ML-250



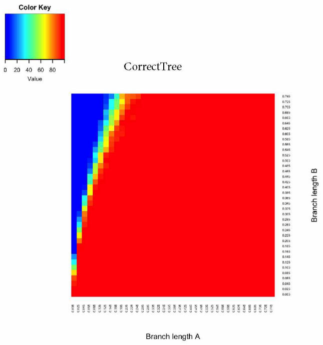
Parsimony-500



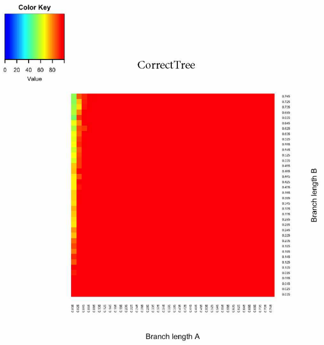
NJ-500



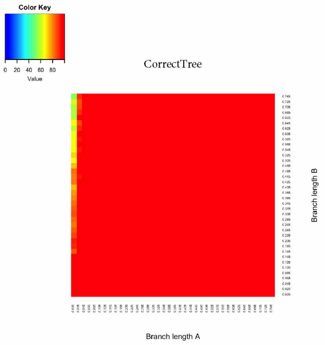
ML-500



Parsimony-1000



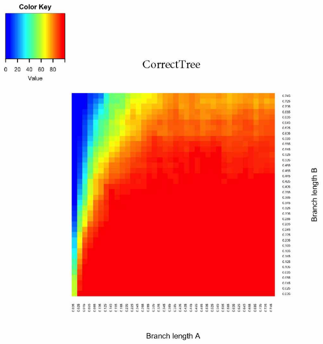
NJ-1000

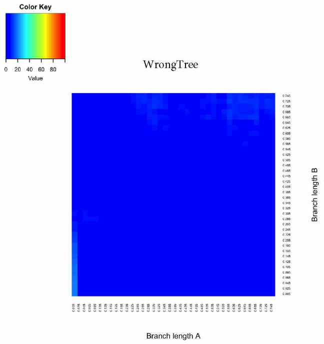


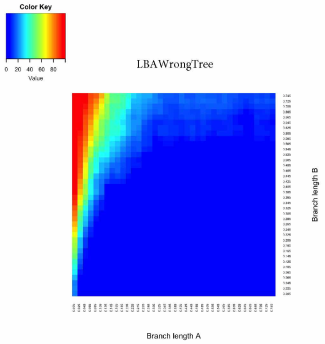
ML-1000

Appendix 6 : Sequences generated under JC and reconstructed under Parsimony

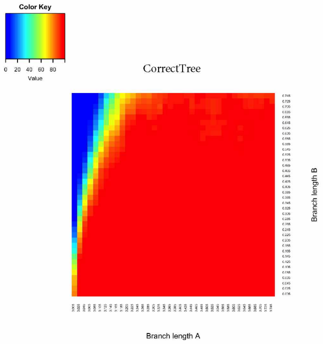
Sequence length 100

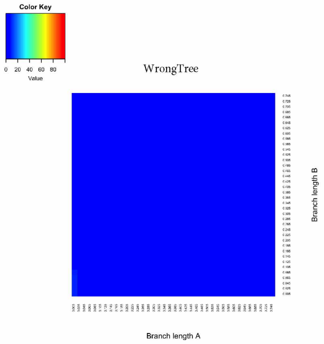


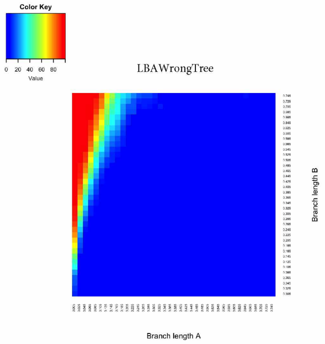




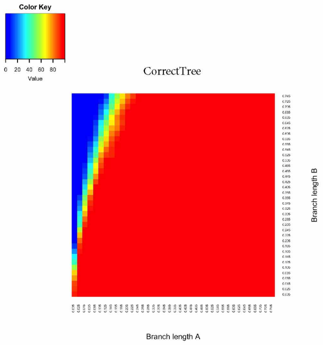
Sequence length 250

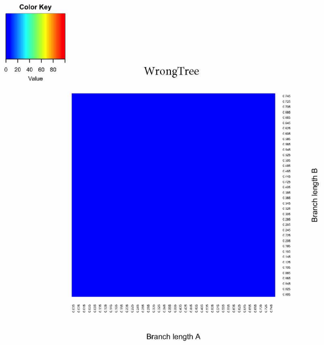


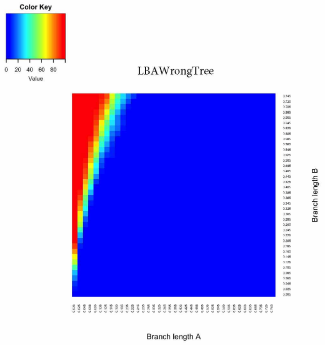




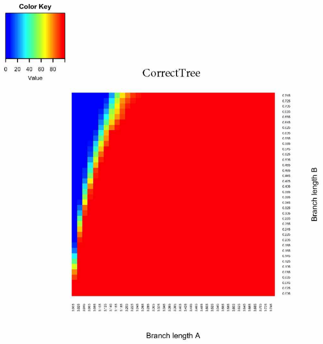
Sequence length 500

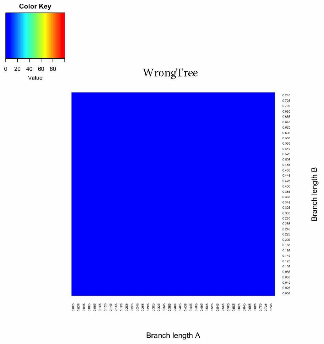


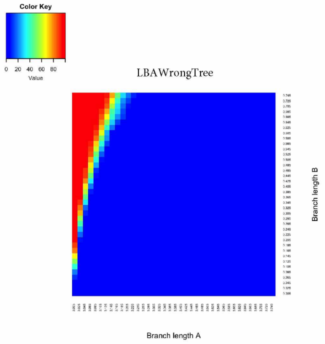




Sequence length 1000

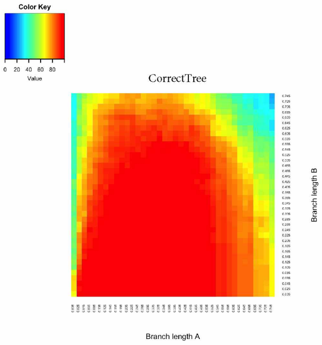


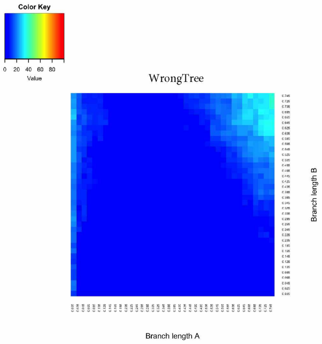


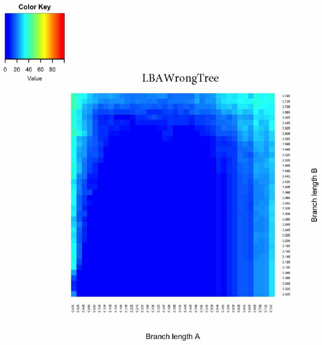


Appendix 7 : Sequences generated under JC and reconstructed under NJ

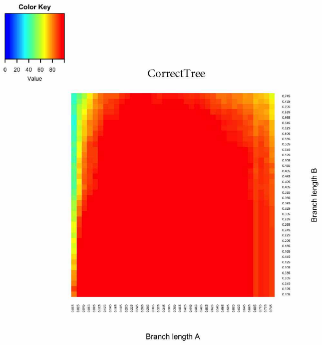
Sequence length 100

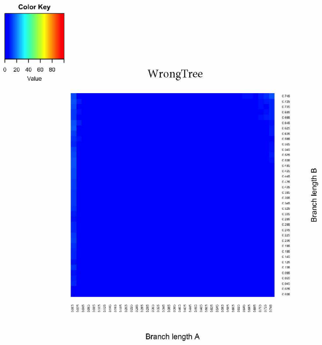


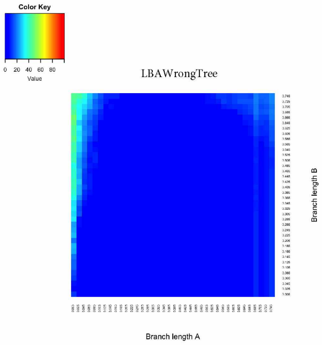




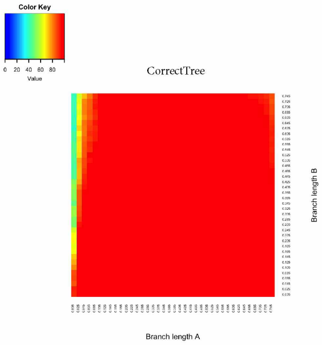
Sequence length 250

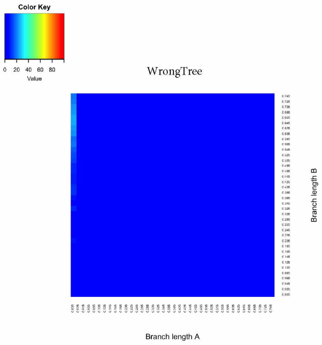


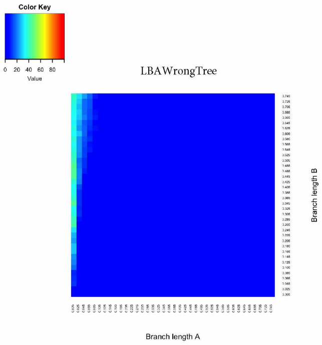




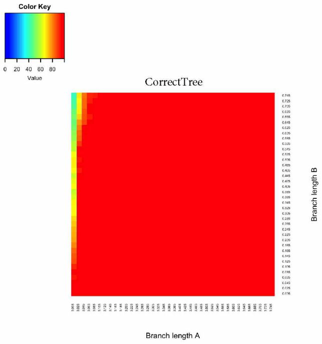
Sequence length 500

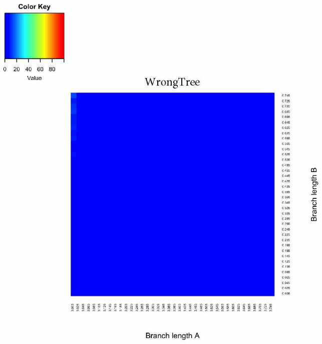


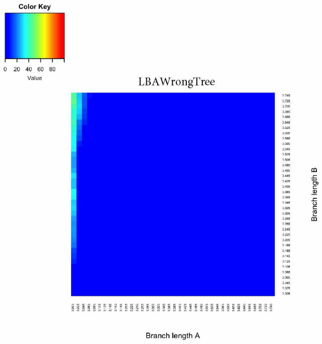




Sequence length 1000

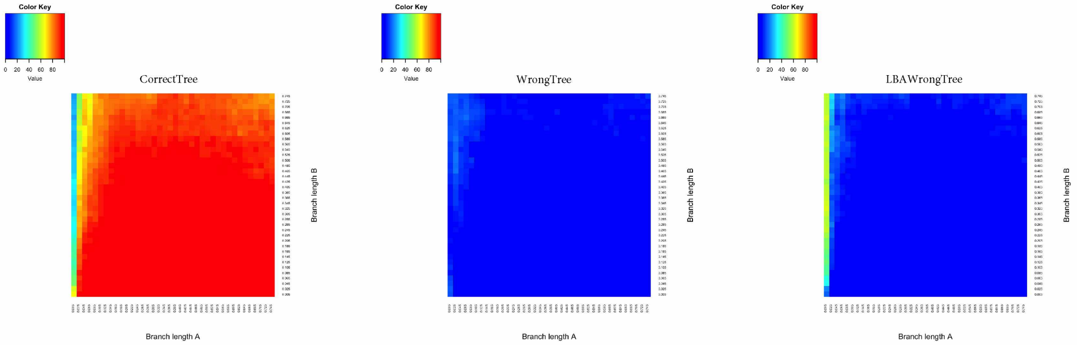




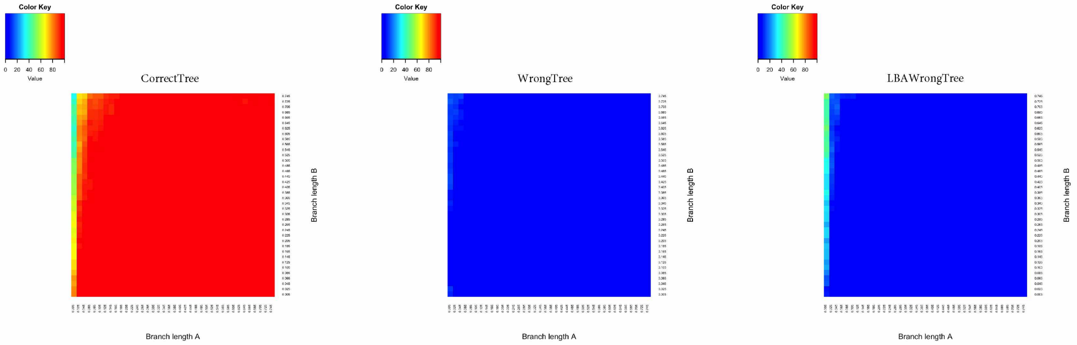


Appendix 8 : Sequences generated under JC and reconstructed under ML

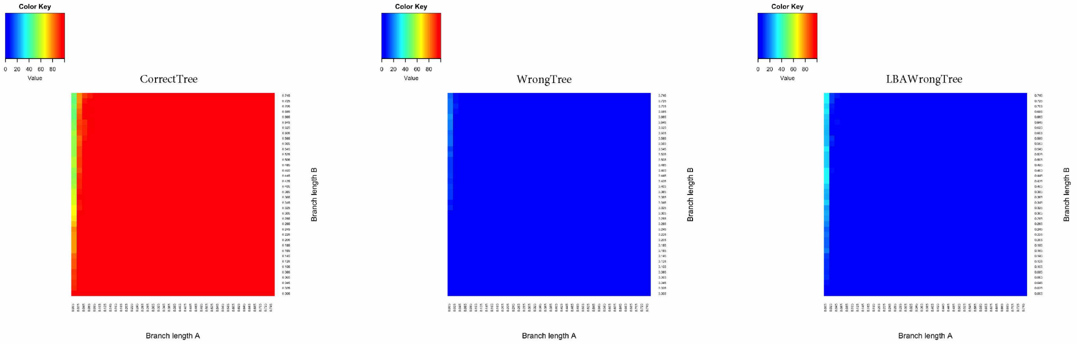
Sequence length 100



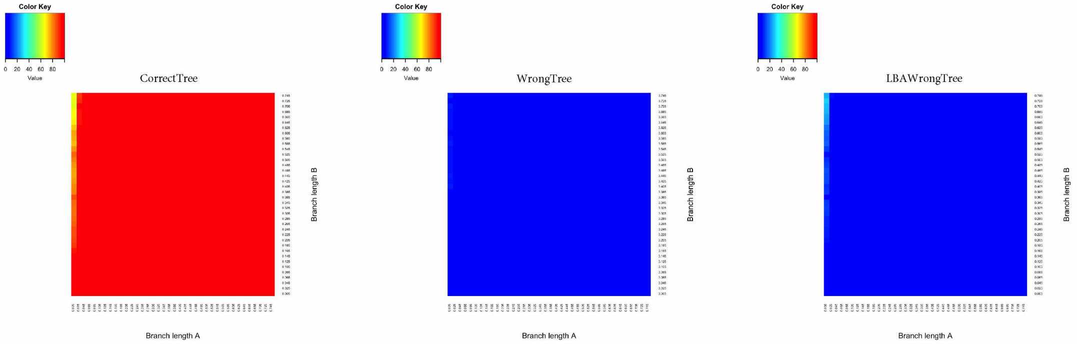
Sequence length 250



Sequence length 500

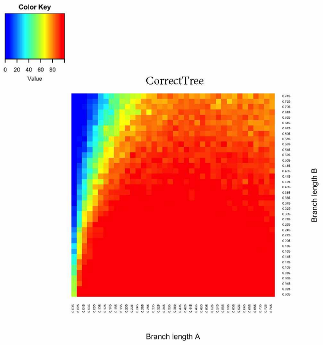


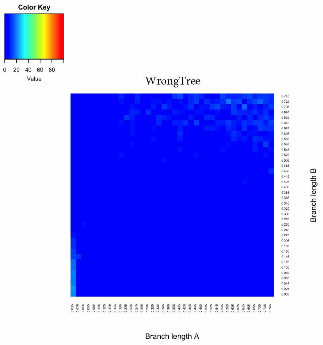
Sequence length 1000

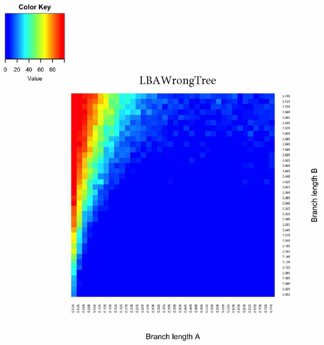


Appendix 9 : Sequences generated under F81 and reconstructed under Parsimony

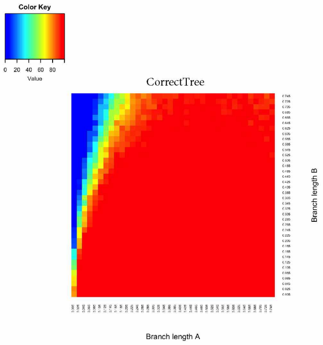
Sequence length 100

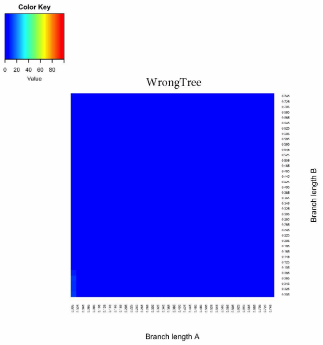


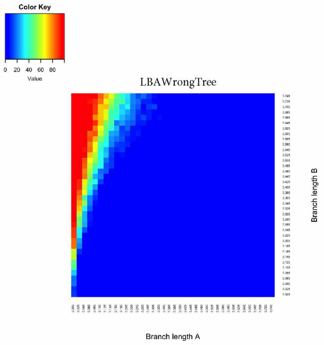




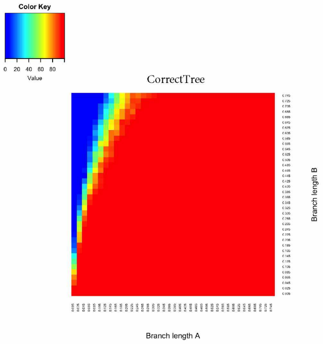
Sequence length 250

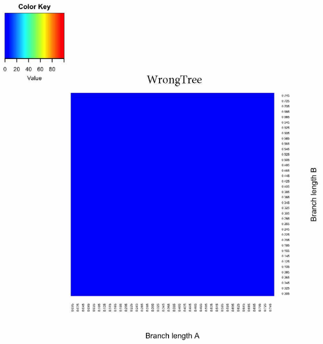


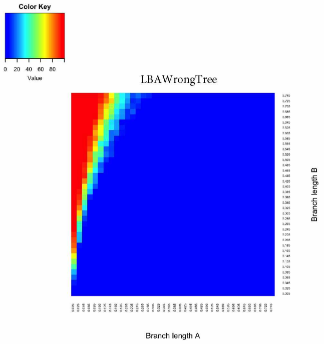




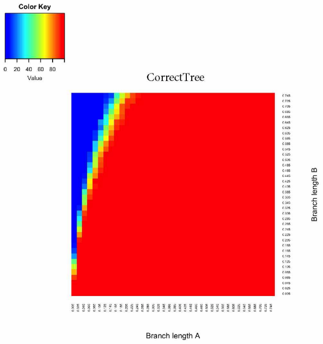
Sequence length 500

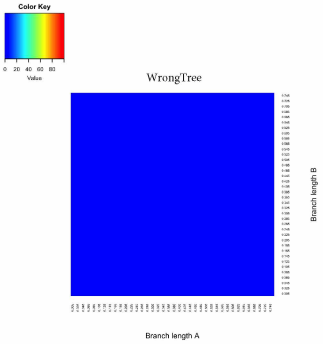


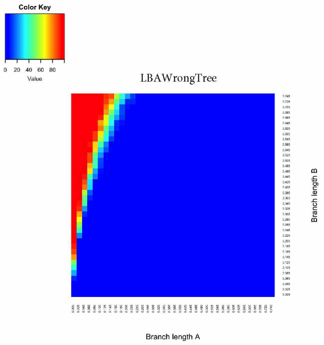




Sequence length 1000

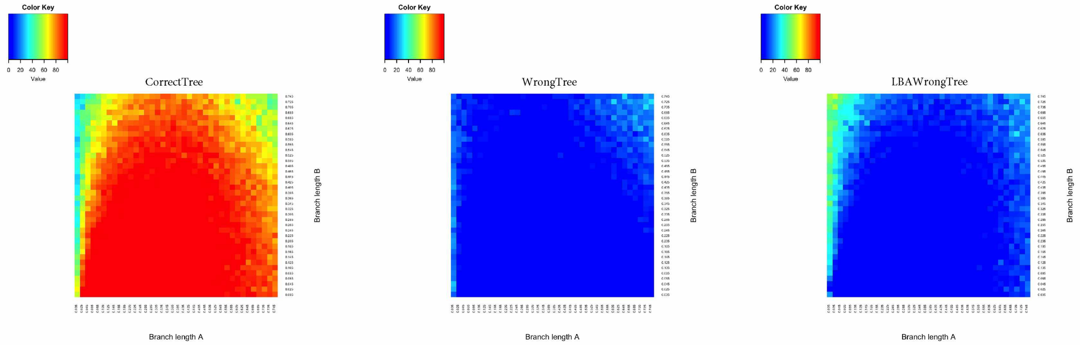




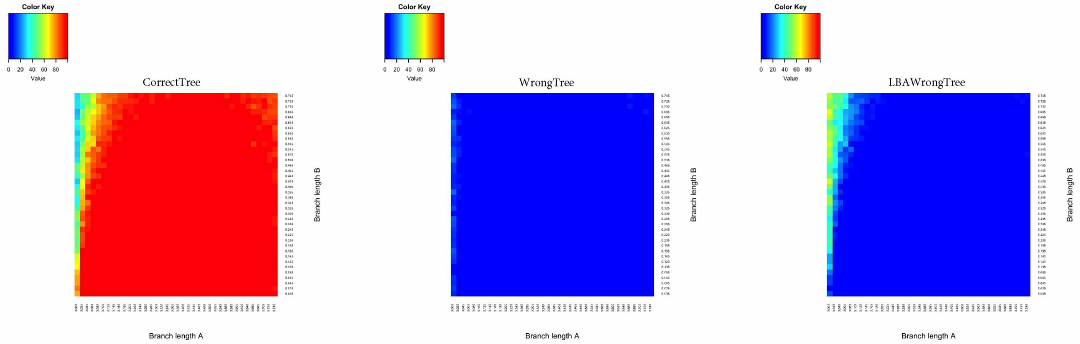


Appendix 10 : Sequences generated under F81 and reconstructed under NJ

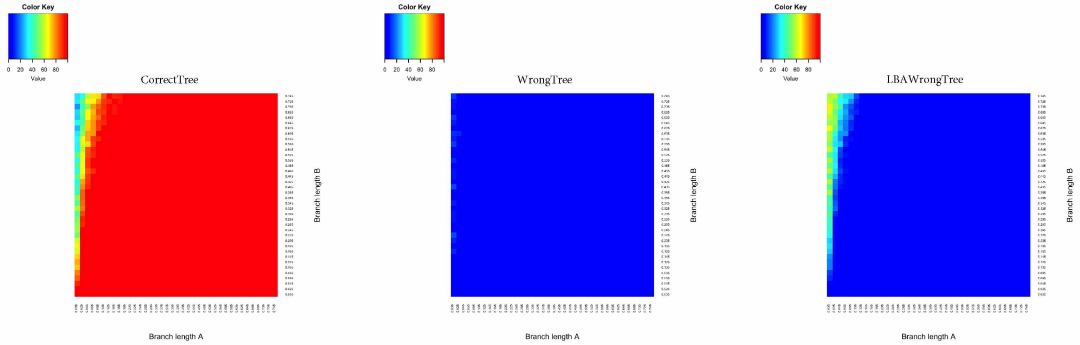
Sequence length 100



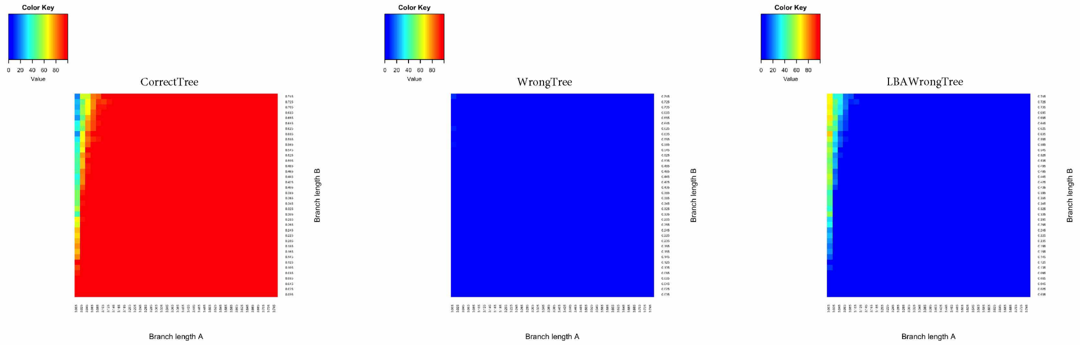
Sequence length 250



Sequence length 500

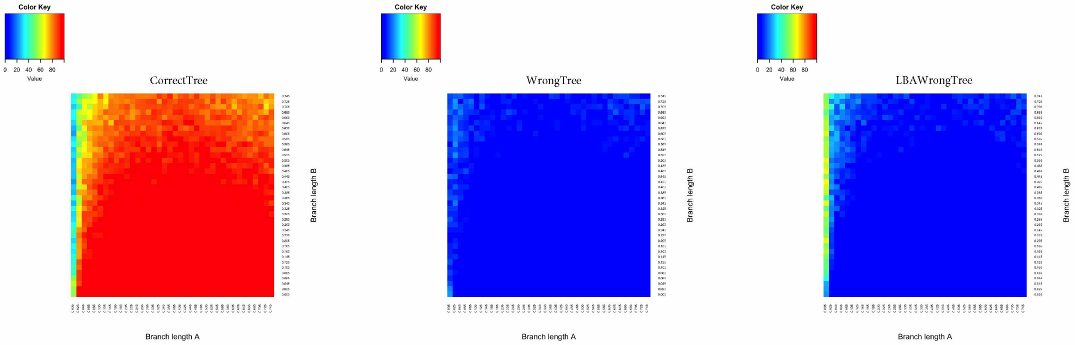


Sequence length 1000

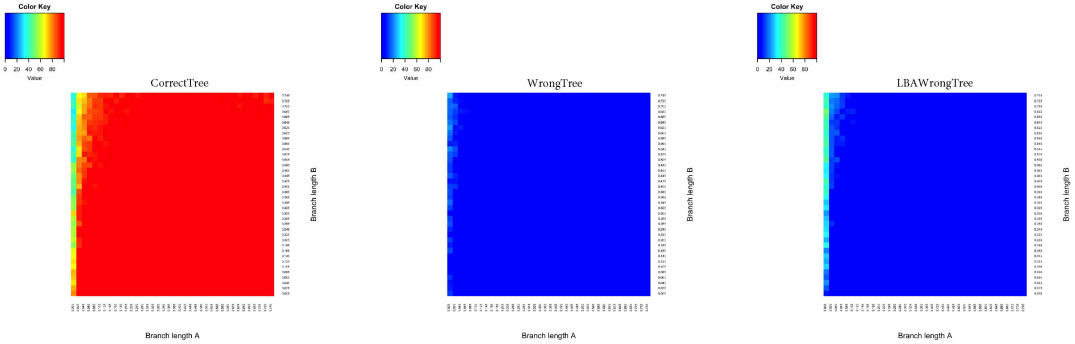


Appendix 11 : Sequences generated under F81 and reconstructed under ML

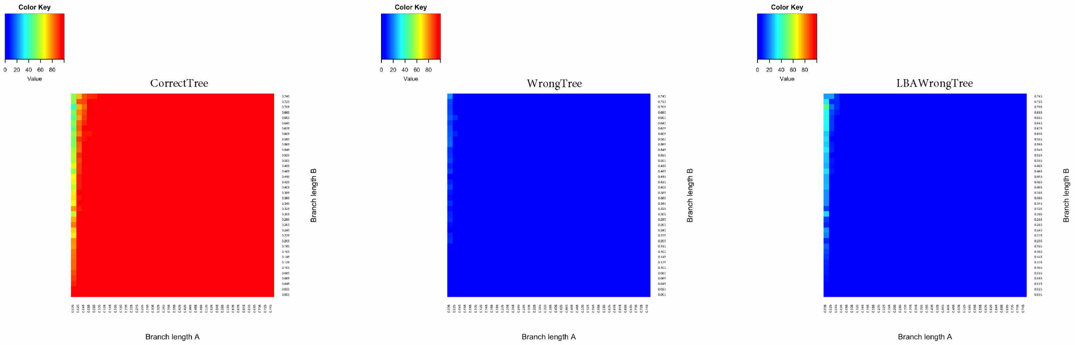
Sequence length 100



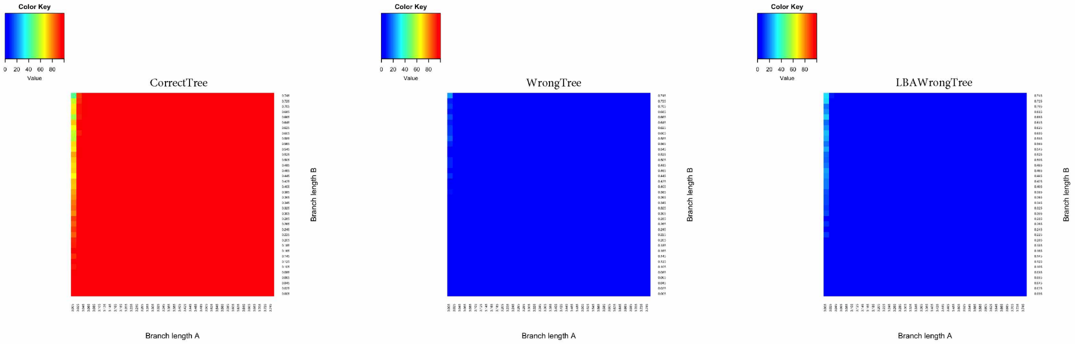
Sequence length 250



Sequence length 500

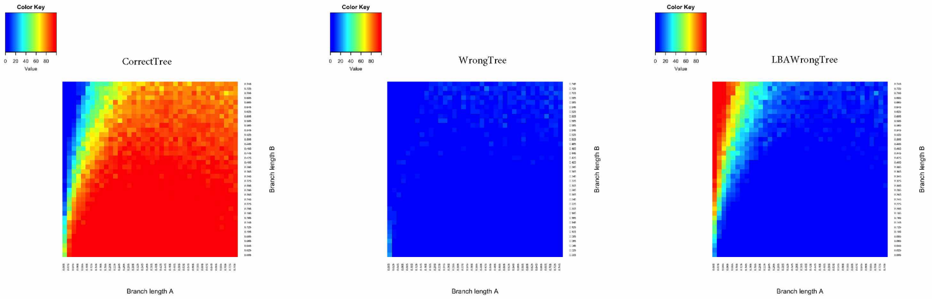


Sequence length 1000

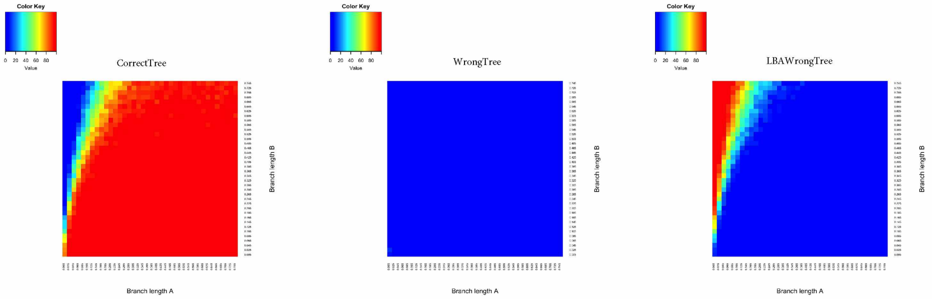


Appendix 12 : Sequences generated under K2P and reconstructed under Parsimony

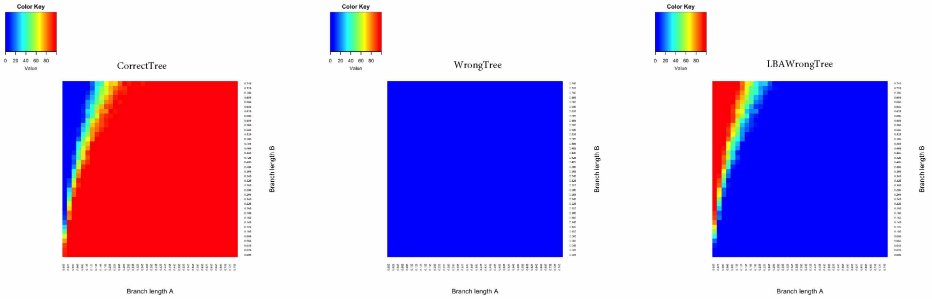
Sequence length 100



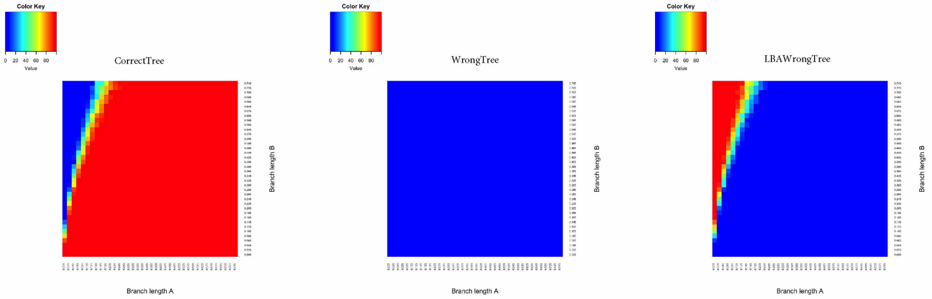
Sequence length 250



Sequence length 500

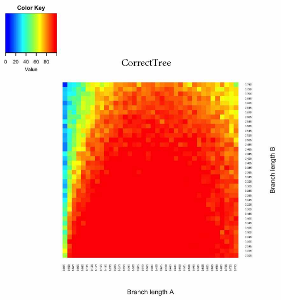


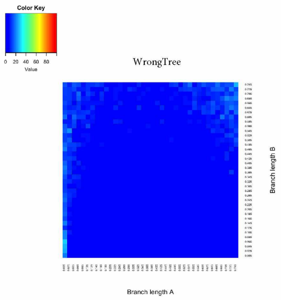
Sequence length 1000

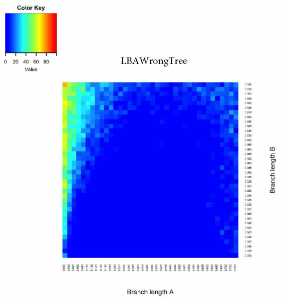


Appendix 13 : Sequences generated under K2P and reconstructed under NJ

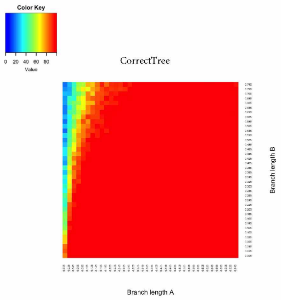
Sequence length 100

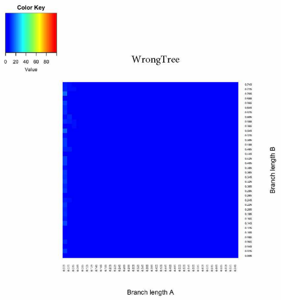


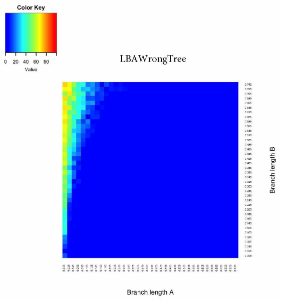




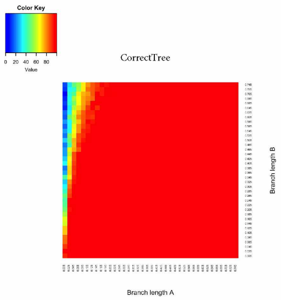
Sequence length 250

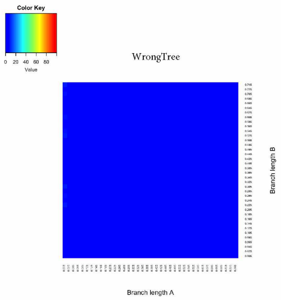


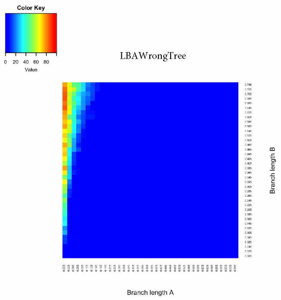




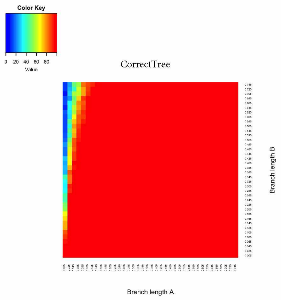
Sequence length 500

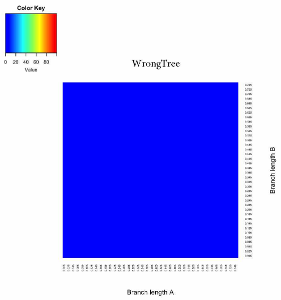


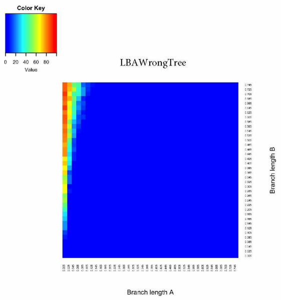




Sequence length 1000

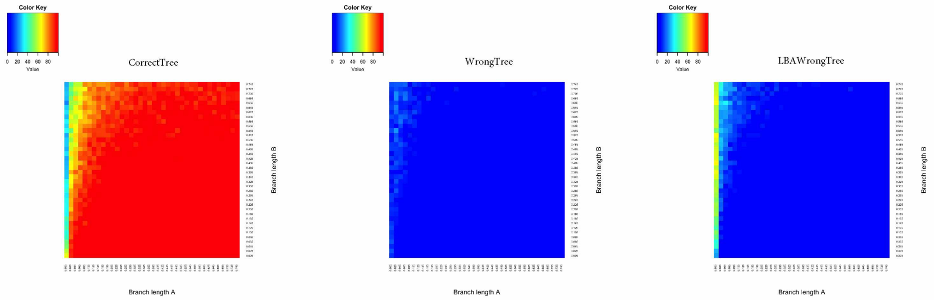




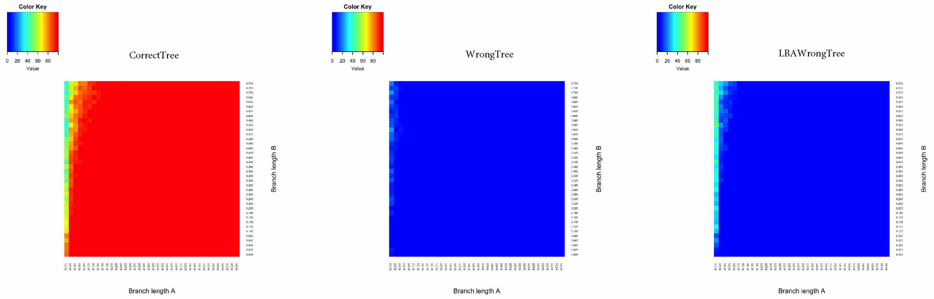


Appendix 14 : Sequences generated under K2P and reconstructed under ML

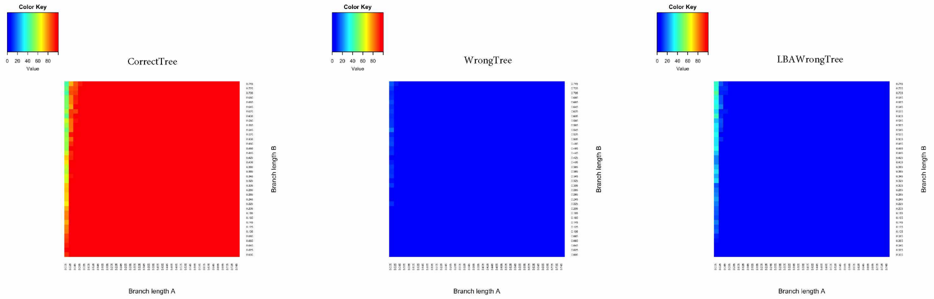
Sequence length 100



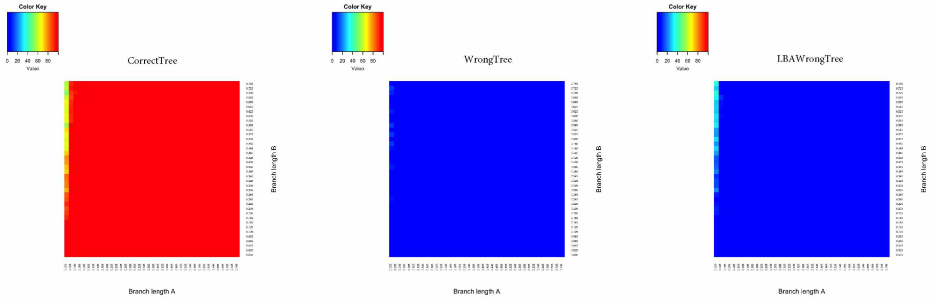
Sequence length 250



Sequence length 500

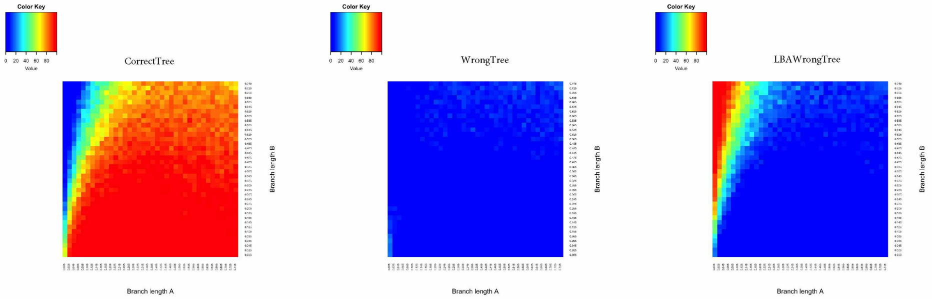


Sequence length 1000

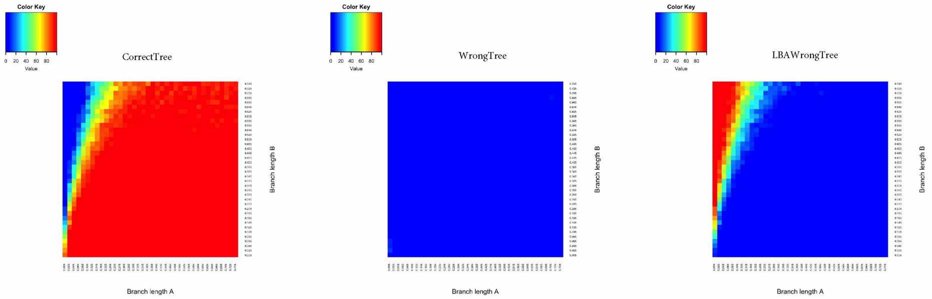


Appendix 15 : Sequences generated under HKY and reconstructed under Parsimony

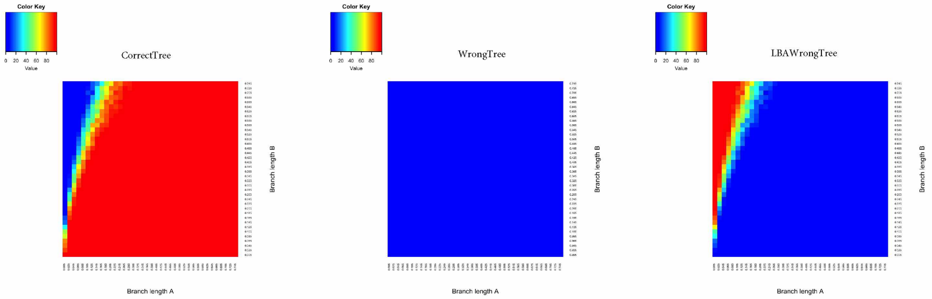
Sequence length 100



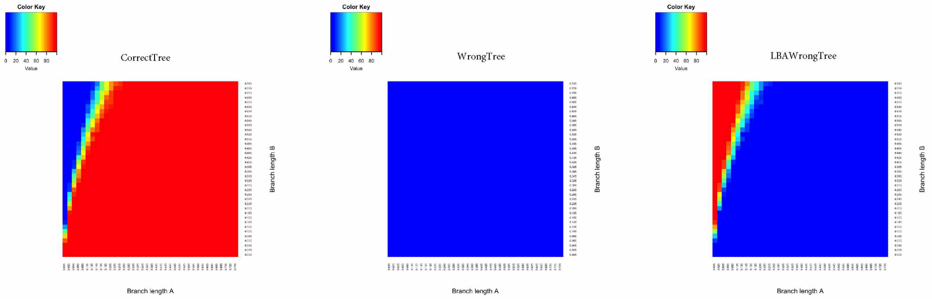
Sequence length 250



Sequence length 500

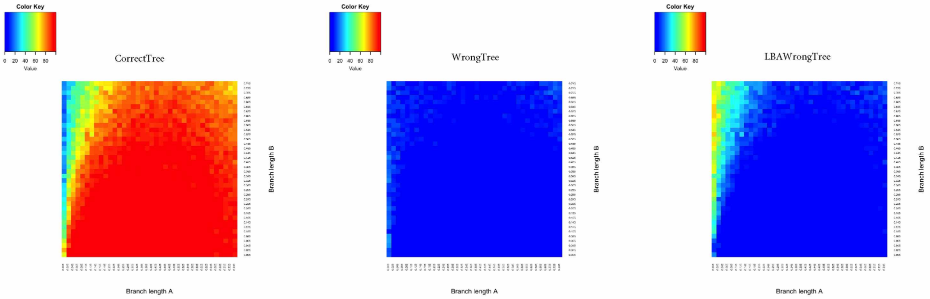


Sequence length 1000

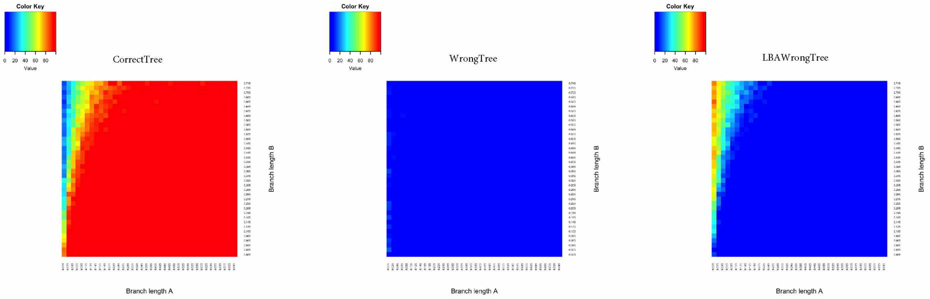


Appendix 16 : Sequences generated under HKY and reconstructed under NJ

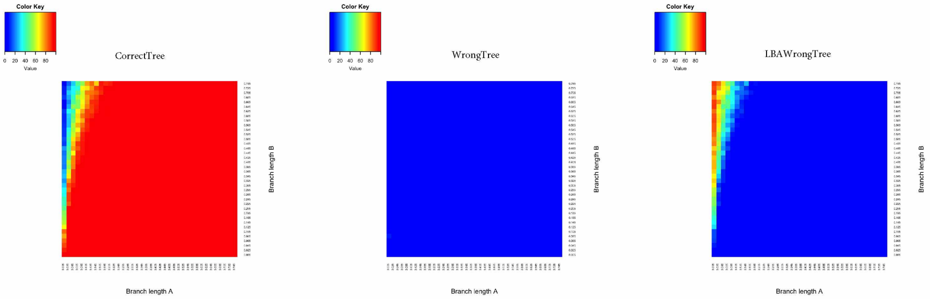
Sequence length 100



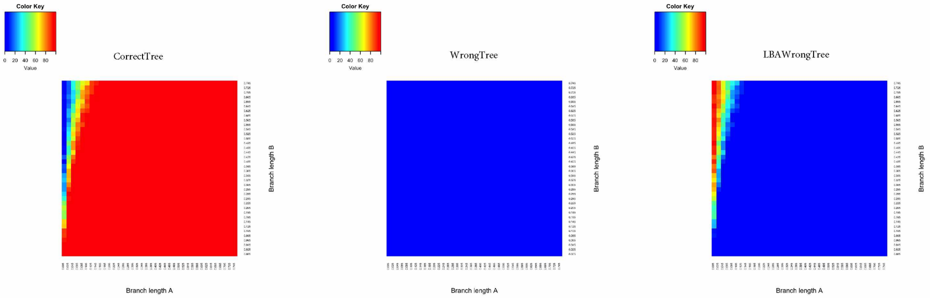
Sequence length 250



Sequence length 500

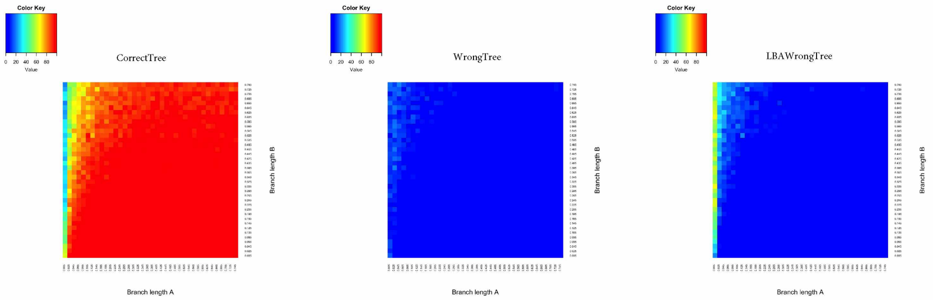


Sequence length 1000

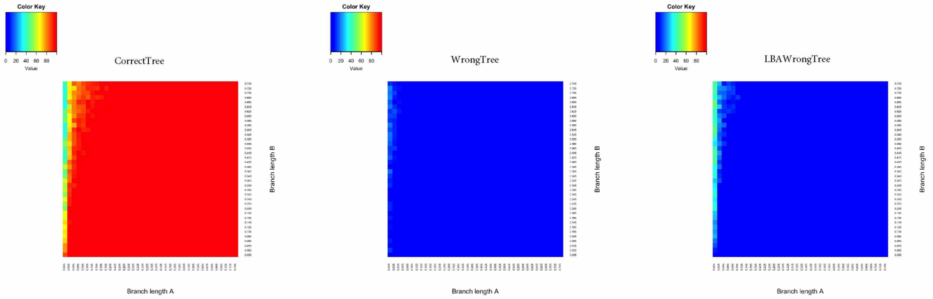


Appendix 17 : Sequences generated under HKY and reconstructed under ML

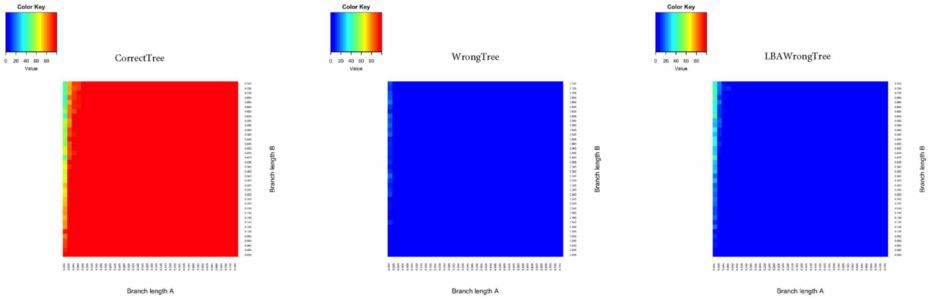
Sequence length 100



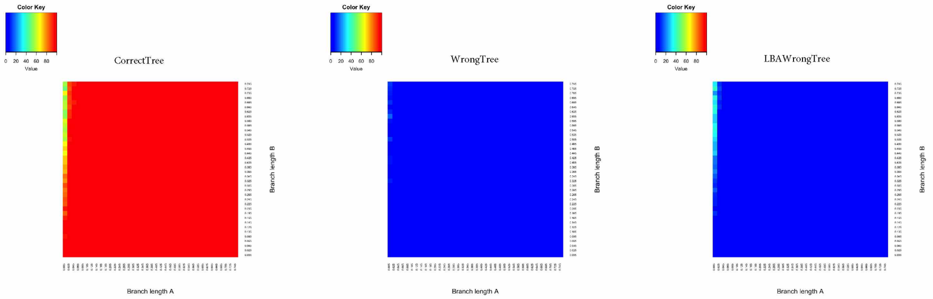
Sequence length 250



Sequence length 500

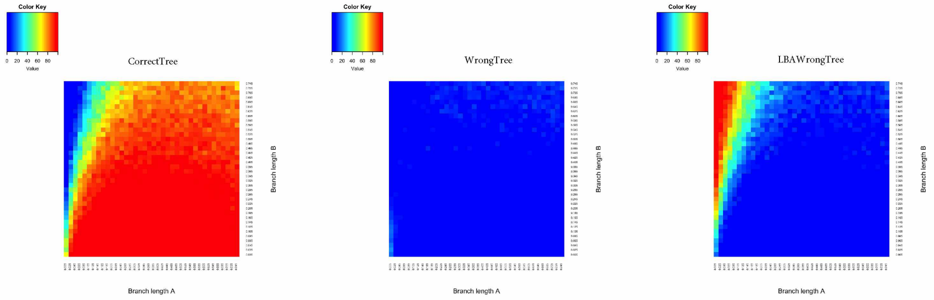


Sequence length 1000

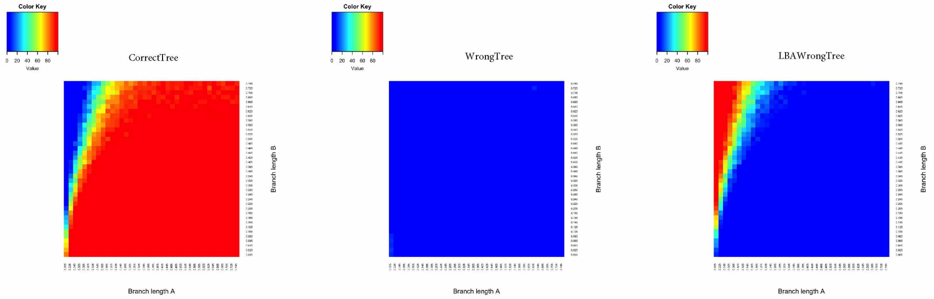


Appendix 18 : Sequences generated under GTR and reconstructed under Parsimony

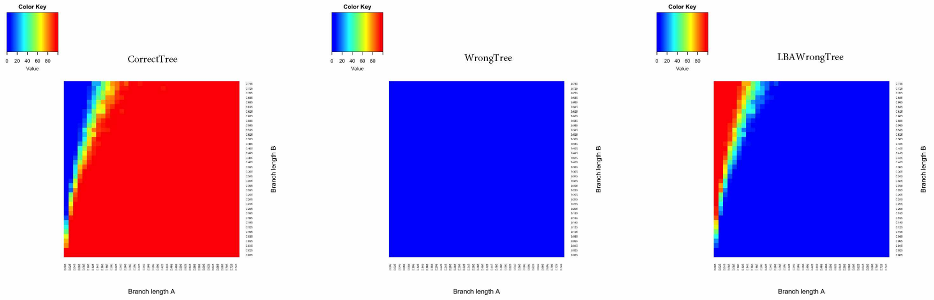
Sequence length 100



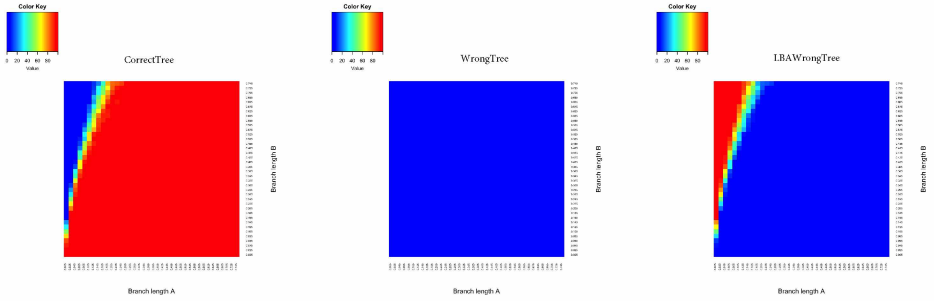
Sequence length 250



Sequence length 500

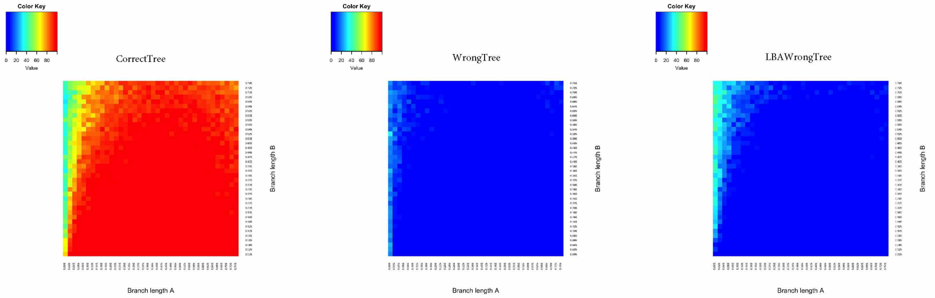


Sequence length 1000

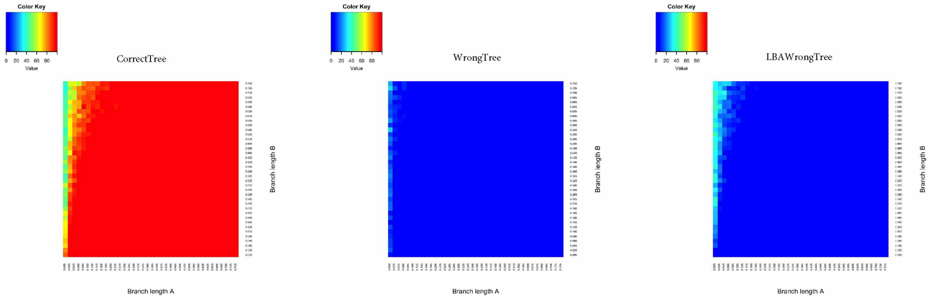


Appendix 19 : Sequences generated under GTR and reconstructed under NJ

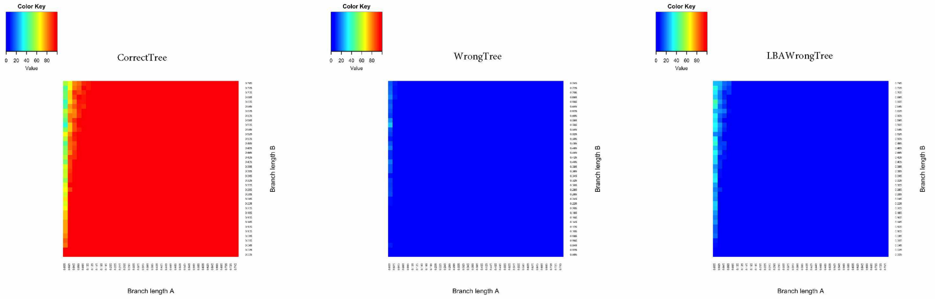
Sequence length 100



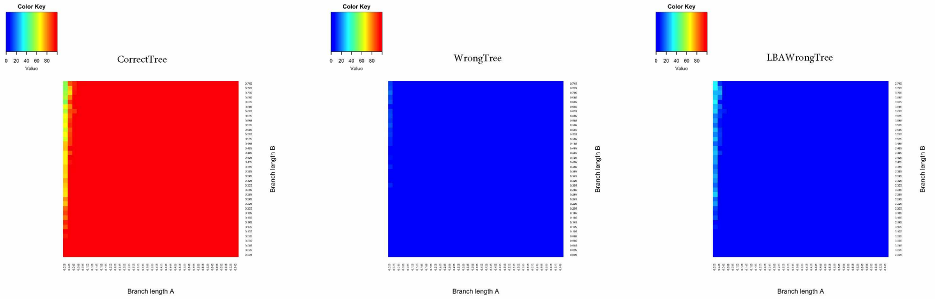
Sequence length 250



Sequence length 500

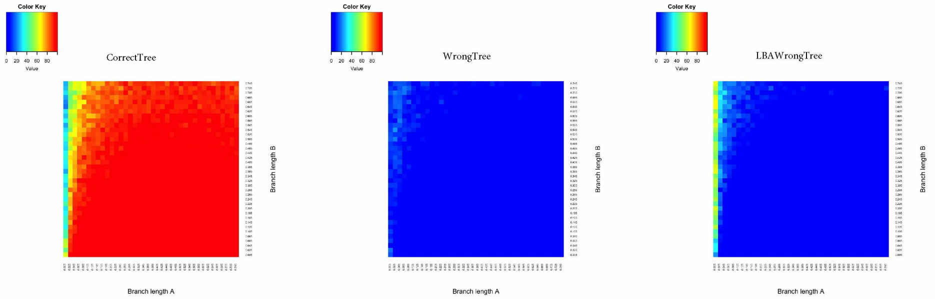


Sequence length 1000

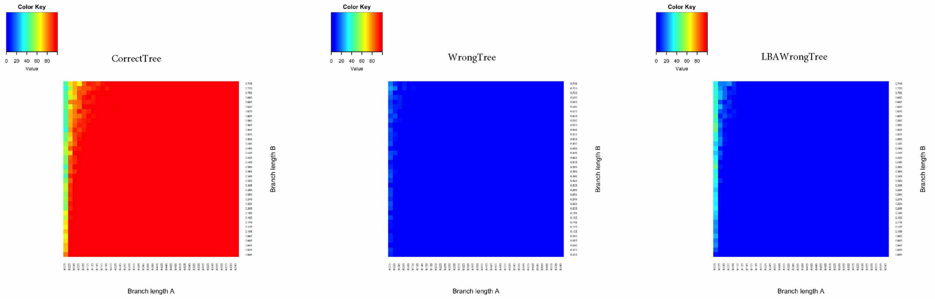


Appendix 20 : Sequences generated under GTR and reconstructed under ML

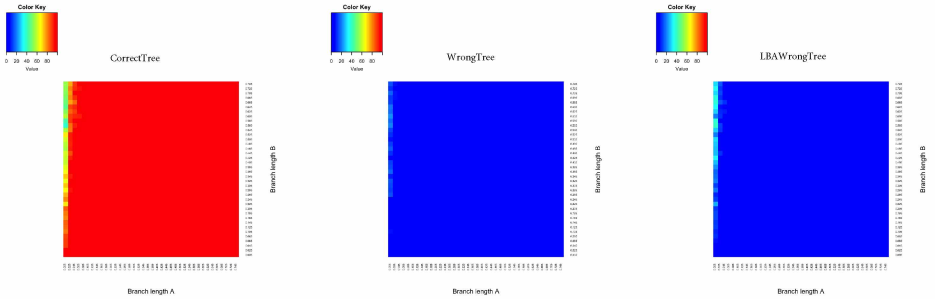
Sequence length 100



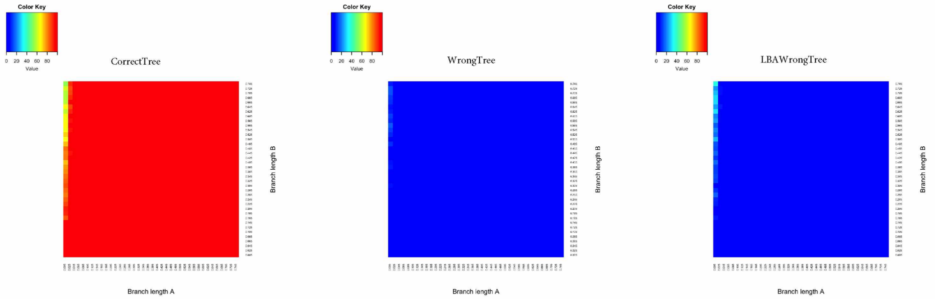
Sequence length 250



Sequence length 500



Sequence length 1000


Color Key
Value
CorrectTree
WrongTree
LBAWrongTree
Branch length B
Branch length A