Rooting Major Cellular Radiations using Statistical Phylogenetics

Svetlana Cherlin

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics & Statistics
Institute for Cell & Molecular Biosciences
Newcastle University
Newcastle upon Tyne
United Kingdom

September 2016

I dedicate this thesis to my husband, Alexander, my daughter, Dana and my son, Idan Itzhak, who kept me going with their support, reassurance and tolerance.

Acknowledgements

I would like to thank my supervisors, Dr. Tom Nye, Prof. Richard Boys and Prof. Martin Embley, for their professional guidance and advice over the course of my postgraduate education. I would like to thank also Dr. Tom Williams for his help with the biological aspects of the thesis. I am immensely grateful to Dr. Sarah Heaps for her patience, kindness, professional and personal help, encouragement and support.

I am grateful to the School of Mathematics & Statistics and to the Institute for Cell & Molecular Biosciences for providing me with all the necessary facilities and resources. I am thankful to Dr. Michael Beaty for his invaluable help with computational issues.

I am also grateful to the European Research Council for the funding which made it possible to conduct this research.

Abstract

Phylogenetics focuses on learning about evolutionary relationships between species. These relationships can be represented by phylogenetic trees, where similar species are grouped together as sharing a recent common ancestor. The common ancestor of all the species of the tree is the root of the tree. The root is fundamental to the biological interpretation of the tree, providing a critical reference point for polarising ancestor-descendant relationships and determining the order in which key traits evolved along the tree (Embley and Martin, 2006). Despite its importance, most models of sequence evolution are unable to infer the root of a phylogenetic tree. They are based on homogeneous continuous time Markov processes (CTMPs) that are assumed to be stationary and time-reversible, with the mathematical consequence that the likelihood of a tree does not depend on where it is rooted. As a result, the root of the tree cannot be inferred as part of the analysis. Other methods which are generally used to root evolutionary trees can be problematic. For example, the outgroup rooting method is susceptible to a long-branch attraction artefact. Paralogue rooting requires pairs of paralogous genes which underwent an ancient gene duplication event to be present in all species being analysed, and the number of such genes is limited.

In this thesis we explore an alternative model-based approach, adopting a substitution model in which changing the root position changes the likelihood of the tree. We explore the effect of relaxing reversibility and stationarity assumptions and allowing the position of the root to be another unknown quantity in the model. We propose two hierarchical non-reversible models which are centred on a reversible model but perturbed to allow non-reversibility. The models differ in the degree of structure imposed on the perturbations. We also explore non-stationary models, and the combination of relaxing both the reversibility and the stationarity assumptions.

The analysis is performed in the Bayesian framework using Markov chain Monte Carlo methods. We illustrate the performance of the models in analyses of simulated datasets using two types of topological priors. We also investigate the effect of different topologies and branch lengths on the inference. Our results illustrate the usefulness of modelling non-reversibility and non-stationarity for root inference, and also demonstrate the sensitivity of the analysis to topological priors. We then apply the models to real biological datasets, the radiation of polyploid yeasts and the radiation of primates, for which there is a robust biological opinion about the root position. Finally we apply the models to an open question in biology: rooting the ribosomal tree of life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

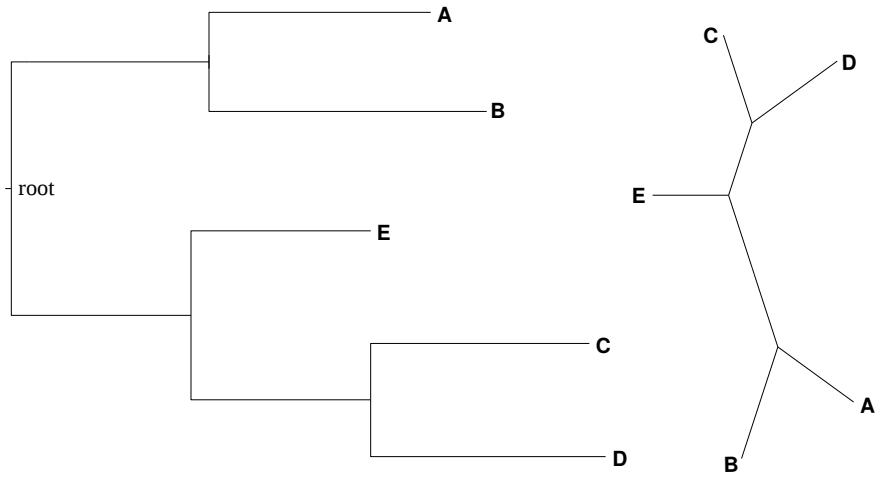
1.1 Molecular phylogenetics

The aim of molecular phylogenetics is to learn about the evolutionary relationships amongst a collection of species using protein or DNA sequences. The main assumption of phylogenetics is that the evolution of life on earth can by represented in the form of a bifurcating tree (*phylogenetic tree*, or *phylogeny*). A tree is a connected acyclic graph with the *leaves* (tips, or external vertices) of the tree representing the extant species. The number of edges connected to a vertex is called the *degree* of the vertex. In a bifurcating tree, all the internal nodes have degree 3 (apart from the root vertex which has degree 2), while the leaves have degree 1. Each edge in the tree represents the period of time over which point mutations accumulate and each bifurcation (vertex) represents a speciation event. The branching pattern of a tree is called the *topology*.

Trees might be either rooted or unrooted. A rooted phylogenetic tree has a special vertex which is denoted the *root* and it represents the most recent common ancestor (MRCA) of all species in the tree. Unrooted trees lack any information about ancestry between the vertices (Figure 1.1). The trees are reconstructed from the alignment of homologous sequences (sequences related to each other by a common ancestor). We assume that each column of the alignment (a site of the alignment) has originated from the same nucleotide of the MRCA (Figure 1.2). A group of species that evolved from the same most recent ancestor is called a *monophyletic group*, or a *clade*. A group of species that evolved from different most recent ancestors is called a *paraphyletic group* (Figure 1.3).

1.2 Motivating example: the tree of life

The *tree of life*, or the *universal tree* describes the evolutionary relationships between all living organisms. The early attempts to infer the tree of life based on molecular data clearly showed three distinct clusters corresponding to the Bacteria, the Archaea and the



(a) Rooted tree. (b) Unrooted tree.

Figure 1.1: Rooted and unrooted 5-species trees. The tree on panel (a) represents a rooted version of the unrooted tree on panel (b).

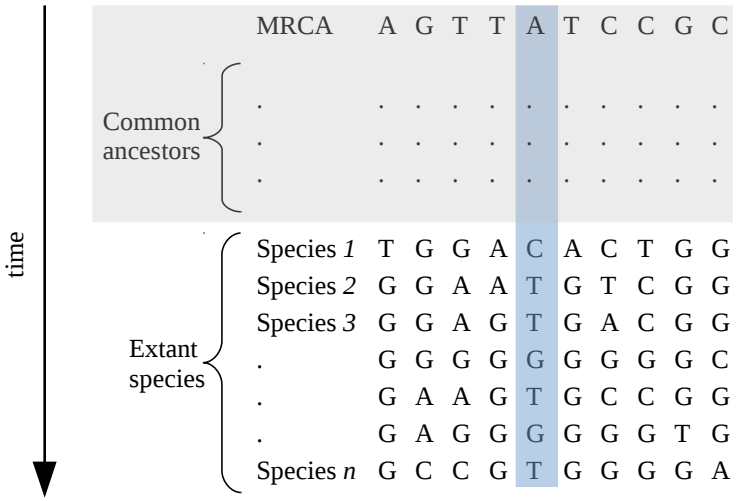


Figure 1.2: An illustration of the known aligned sequences of n extant species and the unknown sequences of their hypothetical common ancestors. Each site evolved from a single nucleotide of the most recent common ancestor (MRCA). The blue region illustrates a single site which has evolved from the nucleotide A of the MRCA. The grey region represents the common ancestors including the MRCA whose unobserved sequences are inferred during the analysis.

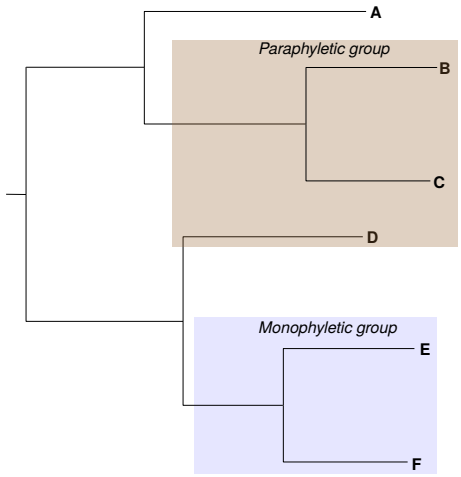


Figure 1.3: An illustration of a monophyletic group (clade) and a paraphyletic group on a phylogenetic tree.

Eukaryota (eukaryotes). These three groups appeared to have branched from the common ancestor at about the same time (Woese & Fox, 1977). The Archaea was subdivided into two domains: Euryarchaeota (encompassing the methanogens and their phenotypically diverse relatives) and Crenarchaeota, also called eocytes (comprising extremely thermophilic Archaea). The analysis of small subunit ribosomal RNA molecules showed that the eukaryotes and the Archaea is a monophyletic group, suggesting a three domains structure of the tree of life (Woese, 1990). However, the structural similarity in ribosomes between the eocytes and the eukaryotes suggested that these two groups are more closely related to each other than to the other Archaea (Lake *et al.*, 1984). This finding gave rise to the eocyte hypothesis whereby the eukaryotic lineage has originated from within a paraphyletic Archaea as a sister group to the eocytes. In its modern formulation the eocyte hypothesis implies that the closest relatives of the eukaryotes are the TACK superphylum which includes recently discovered relatives of the eocytes (Guy & Ettema, 2011; Kelly *et al.*, 2011; Williams *et al.*, 2013) (Figure 1.4). Even though the three-domains hypothesis is the dominant paradigm, there is increasing support for the eocyte hypothesis from recent published studies (Embley & Martin, 2006; Cox *et al.*, 2008; Williams *et al.*, 2012, 2013; Heaps *et al.*, 2014; Spang *et al.*, 2015). The root of the tree of life is also a highly debated issue in biology. While widely agreed opinion places the root on the branch leading to the Bacteria, a few studies have suggested that the root is within the Bacteria (Lake *et al.*, 2009; Skophammer *et al.*, 2007; Heaps *et al.*, 2014), or within the eukaryotes (Brinkmann & Philippe, 1999; Philippe & Forterre, 1999).

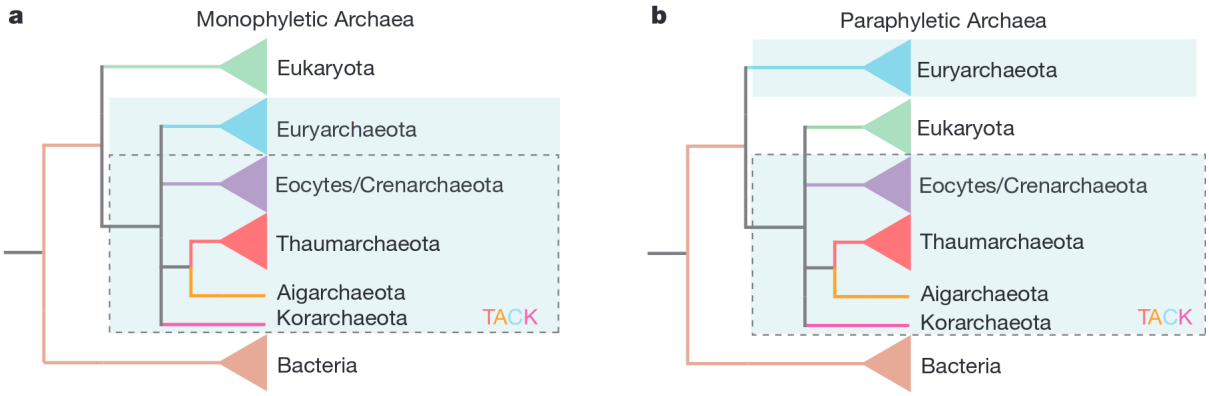


Figure 1.4: Two competing hypothesis about the tree of life, from Williams *et al.* (2013). Panel (a) depicts a three-domains tree of life, where monophyletic Archaea (blue background) share a common ancestor with the eukaryotes. Panel (b) represent the alternative eocyte hypothesis whereby the eukaryotes have originated from within the paraphyletic Archaea (blue background) and are more closely related to the TACK superphylum.

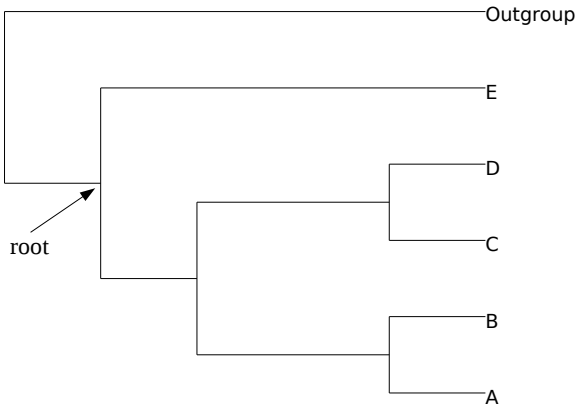
1.3 Rooting a phylogenetic tree

A root of a phylogenetic tree is a key component of phylogenetic inference, providing a point of reference for investigating fundamental biological questions about the evolution of species, such as polarising ancestor/descendant relationships and ancestral state reconstruction. In this section we will review current methods used to root phylogenetic trees.

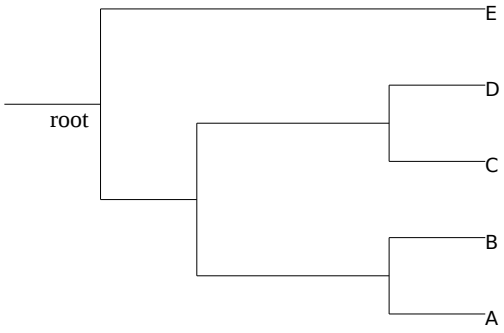
1.3.1 Outgroup rooting

This method uses an *outgroup* which is one or more taxa known to lie outside of the clade for which the root is being investigated (the *ingroup*). According to this method, an unrooted tree for a data set comprising sequences from both the ingroup and the outgroup is constructed. The branch connecting the ingroup and the outgroup becomes the root of the tree for the species of interest (Figure 1.5) (Penny, 1976; Huelsenbeck *et al.*, 2002). For example, Cannarozzi *et al.* (2007) used the opossum as an outgroup in order to determine the evolutionary relationships between human, dog and mouse (Figure 1.6).

However, this approach can be problematic if the outgroup is only distantly related to the ingroup, because the long branch leading to the outgroup can induce phylogenetic artefacts such as long branch attraction (LBA), whereby long branches tend to group together on a tree irrespective of their true evolutionary relationships. Thus, the long branch leading to the outgroup can potentially interfere with the inference of ingroup relationships and the root position (Felsenstein, 1978; Holland *et al.*, 2003; Bergsten, 2005). Another drawback of outgroup rooting has been observed when the ingroup and outgroup



(a) Unrooted tree including the ingroup and the outgroup. The root is placed on the branch connecting the ingroup to the outgroup.



(b) Tree rooted by the outgroup method.

Figure 1.5: Outgroup rooting of species A, B, C, D and E.

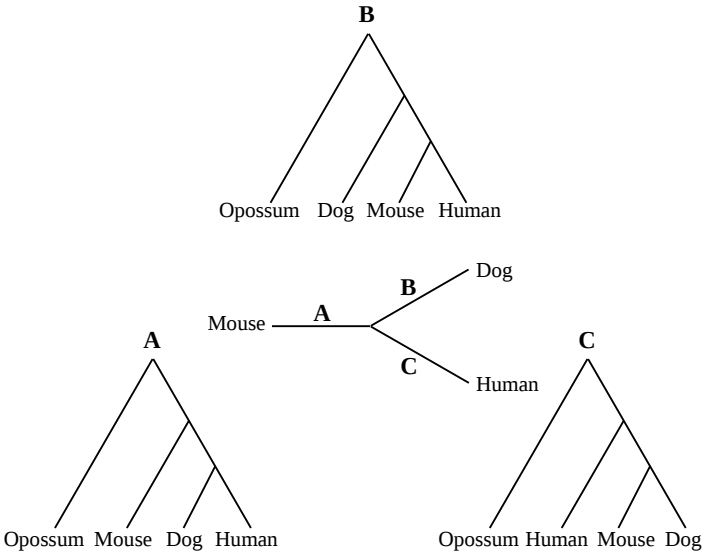
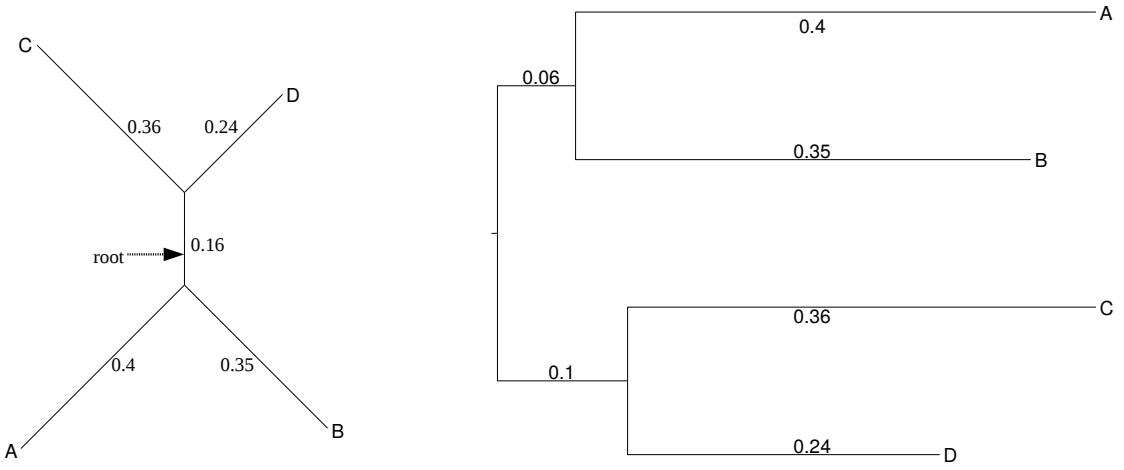


Figure 1.6: An illustration of an outgroup rooting of human, dog and mouse using the opossum as an outgroup (Cannarozzi *et al.*, 2007). The tree in the centre of the figure is an unrooted tree of human, dog and mouse. Trees A, B and C represent three possible evolutionary relationships between human, dog and mouse using opossum as an outgroup.



(a) Unrooted tree.

(b) Rooted tree.

Figure 1.7: Midpoint rooting of a 4-taxon tree. The distance between the species A and C is $0.36 + 0.4 + 0.16 = 0.92$ which is the maximum pairwise distance. Thus the root is placed at the distance $0.96/2 = 0.46$ from both A and C (at the internal node).

taxa differ substantially in nucleotide or amino acid composition: the position of the root of the ingroup becomes unstable, depending on the model used to infer the tree (Tarrío *et al.*, 2000; Foster, 2004). Outgroup rooting is also difficult to apply to the question of rooting trees of species for which no obvious outgroup is available, for instance for rooting the universal tree (Iwabe *et al.*, 1989; Brown & Doolittle, 1995; Hashimoto & Hasegawa, 1996; Baldauf, 1996).

1.3.2 Midpoint rooting

Midpoint rooting (Farris, 1972) is useful in situations where no outgroups are available (Sanderson & Shaffer, 2002). According to this method, the root is placed at a point on the tree halfway between the two most distant species. The pairwise distances of all the species on the tree are calculated, and the root is placed on the middle of the path connecting the two species having the biggest pairwise distance (Figure 1.7). Midpoint rooting was tested across multiple studies and it has displayed a high success rate of inferring roots of phylogenetic trees (Hess & Russo, 2013). However, this method requires that the most divergent species on the tree evolve at the same rate (Tarrío *et al.*, 2000; Huelsenbeck *et al.*, 2002), and this assumption is often not credible.

1.3.3 Rooting by gene duplication

This method has been used to root the universal tree of life, for which no outgroup species exists. The method makes use of paralogous genes (genes which have originated as a result

of a duplication event). The strategy is to use pairs of paralogous genes which underwent a gene duplication in the last universal common ancestor prior to the divergence of the three domains of life: Archaea, Bacteria and eukaryotes. As a result of the duplication each one of the species has two copies of the gene (paralogues), and the trees constructed from both copies are similar, therefore the trees can act as an outgroup to each other. An unrooted tree from both paralogues is constructed, and then rooted on the branch connecting the two duplicates (Figure 1.8). Paralogue rooting studies of the tree of life were consistent in placing the root of the tree of life on a branch separating the Bacteria from the eukaryotes and the Archaea (Gogarten *et al.*, 1989; Iwabe *et al.*, 1989; Brown & Doolittle, 1995; Hashimoto & Hasegawa, 1996; Baldauf, 1996).

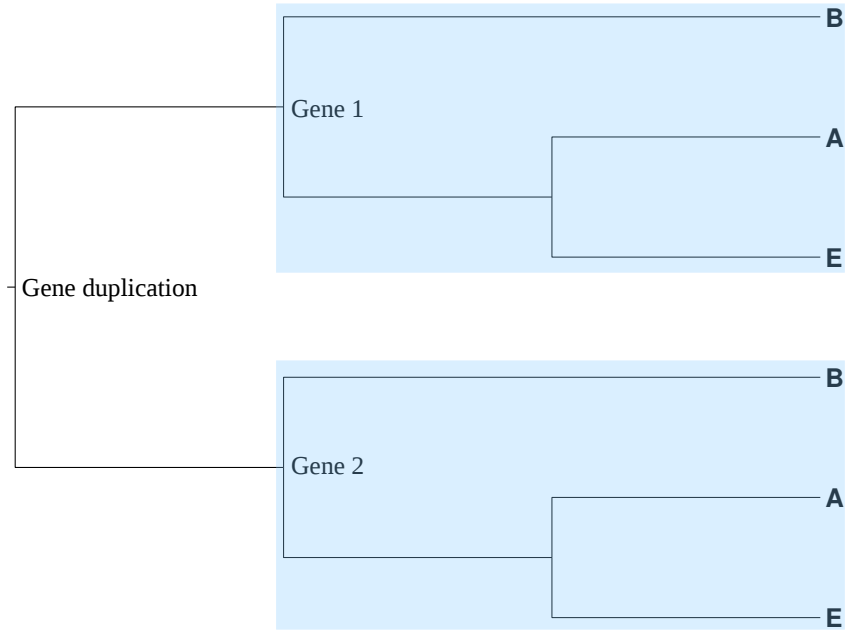


Figure 1.8: Rooting the tree of life by a gene duplication event. Genes 1 and 2 represent a pair of paralogues which originated by a duplication prior to the divergence of the three domains of life: Archaea (A), Bacteria (B) and eukaryotes (E). The tree constructed from both paralogues comprises two similar sub-trees which act as an outgroup to each other, therefore the root is placed on a branch connecting the two sub-trees.

However, the validity of gene duplication rooting has been questioned on various aspects. One of them is detection of anciently duplicated genes. It is difficult to unambiguously establish for any given gene that a duplication took place. The number of genes to which this technique can be applied is also limited. An additional issue is the possibility of artefacts of phylogenetic reconstruction, of which the most important is the long branch attraction (LBA). For example, it has been been suggested that the bacterial rooting of the tree of life is an LBA artefact since the bacterial sequences were determined to have evolved faster than the archaeal and the eukaryotic ones (Philippe & Forterre, 1999;

Zhaxybayeva *et al.*, 2005).

1.3.4 Indel rooting

This method roots trees based on the pattern of *indels*. "Indel" is a term for the insertion or the deletion of nucleotides (one term is used for both phenomena because it is often difficult to establish whether an insertion or a deletion took place). The process of indel rooting works by excluding the root from regions of the tree using a parsimony method (a method that requires the fewest evolutionary changes to explain the differences among the observed sequences). For every possible root position the number of indels needed to produce the observed pattern is calculated. The placement of the root is then defined such that it corresponds to the minimum score of indels.

Indel rooting utilises two paralogous genes. If the indel under analysis is present in only one gene then most parsimoniously the root is excluded from the region containing the indel (Lake *et al.*, 2007). The logic of indel rooting is illustrated in Figure 1.9. The top two sequences of Gene 1 contain an insertion (highlighted), whereas the bottom three sequences of Gene 1 and all the sequences of Gene 2 lack the insertion. The tree on the right side of the figure is rooted within the highlighted clade, while the tree on the left side is rooted outside of the highlighted clade. The rooting of the tree on the left side of the figure requires only an insertion to produce the observed indel pattern. The rooting of the tree on the right side of the figure requires two changes: an insertion somewhere between the two genes, and a deletion on the branch leading from the highlighted clade to the other species on the tree for Gene 1. Thus, the root of the illustrated tree is most parsimoniously placed outside of the highlighted region (root 1) (Lake *et al.*, 2009).

Interestingly, the result of indel based rooting of the universal tree disagrees with the paralogue rooting result. Indel analyses of different proteins excluded all positions of the root from the tree of life except for the branch between actinobacteria and clostridia, thus placing the root within the Bacteria (Lake *et al.*, 2009). Another indel analysis also supports the bacterial rooting (Skophammer *et al.*, 2007), however, the support is given to a few different positions of the root within the Bacteria.

1.3.5 Molecular clock rooting

Molecular clock rooting is based on a molecular clock assumption which asserts that the rate of sequence evolution is constant over time. Under this assumption the expected distance between sequences increases linearly with their time of divergence (more technical explanation will be provided later in this section). However, the assumption of a single constant molecular clock is too simplistic, because the rates of molecular evolution can vary significantly. For example, it has been found that the rate of molecular evolu-

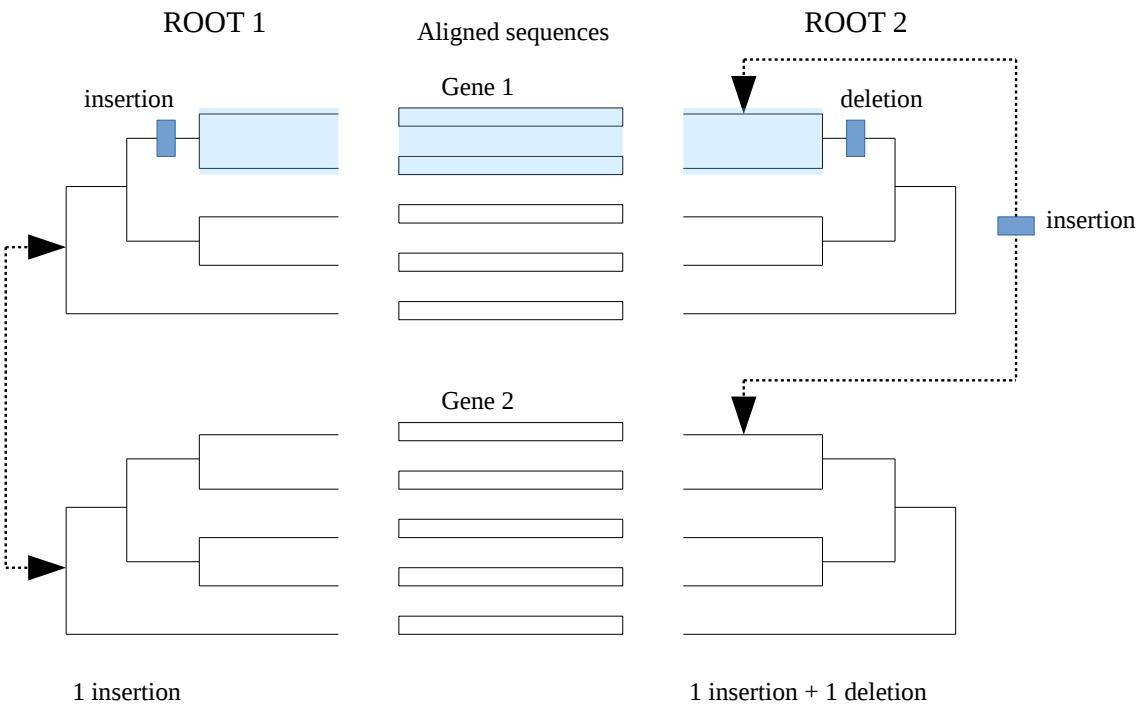


Figure 1.9: Process of indel rooting illustrated for two alternative rootings, using two paralogous genes. In the centre of the figure, the top two sequences of Gene 1 contain an insertion (highlighted), whereas the bottom three sequences of Gene 1 and all the sequences of Gene 2 lack the insertion. The trees on the left and the right sides of the figure represent two different rooted trees that relate the sequences. The tree on the right is rooted through the highlighted region corresponding to those sequences that contain the insert, and the tree on the left is rooted outside of the highlighted region. The right tree is less parsimonious than the left tree, indicating that the root of the tree cannot be placed within the highlighted region (Lake *et al.*, 2009).

tion between the *ribulose-1,5-bisphosphate carboxylase (rbcL)* gene sequences among seed plants (Bosquet *et al.*, 1992), between a variety of mitochondrial and nuclear genes in mammalian lineages (Bromham *et al.*, 1996) and birds (Mooers & Harvey, 1994) exhibit substantial variation. The performance of this method has been shown to deteriorate as the substitution process deviated from the clock assumption (Huelsenbeck *et al.*, 2002). There are also methods that relax the clock assumption allowing for limited variation of the rates of molecular evolution. Relaxed-clock-rooting has been found to be able to correctly identify the root of the tree even when the clock criterion was violated (Renner, 2008), and it has therefore been suggested that relaxed clock rooting can be used even when the substitution process is not strictly clock-like (Renner, 2008; Huelsenbeck *et al.*, 2002). However, the molecular clock assumptions might not be appropriate for distantly related species because their overall rates of molecular evolution might evolve over time (Kumar, 2005).

1.3.6 Model-based approaches

An alternative, but perhaps under-explored, approach to rooting trees is to take a model-based approach, adopting a substitution model in which changing the root position changes the likelihood of the tree. Focusing on homogeneous CTMPs, it is helpful to distinguish between the ideas of *stationarity*, *reversibility* and *homogeneity*. We say that a model is *homogeneous* if it can be characterised by a single instantaneous rate matrix which applies to the whole tree. A homogeneous model is termed *reversible* if the rate matrix can be factorised into a symmetric matrix of exchangeability parameters and a diagonal matrix of stationary probabilities. Similarly we call a rate matrix *reversible* if it permits such a factorisation. Finally a CTMP is *stationary* if the probability of being in each state (e.g. each nucleotide for DNA) does not change over time and the probabilities of transitioning between states over some time interval depend only on the size of that interval and not on its position in time. It follows that all non-stationary models are also non-homogeneous, although the converse need not be true. Models in which one or more of these assumptions is violated can give rise to likelihood functions that depend on the position of the root.

For most models that allow root inference, the focus has been on relaxing the assumption of homogeneity, typically assigning different reversible rate matrices to different parts of the tree. Generally, these models are non-stationary and allow variation in the theoretical stationary distribution across the tree. Some also allow variation in the exchangeability parameters (Dutheil & Boussau, 2008) although these are often fixed over all branches. For example, Yang & Roberts (1995) assigned common exchangeabilities but a different composition vector to each edge of the tree. Heaps *et al.* (2014) fitted a similar model in a Bayesian framework, but adopted a prior over composition vectors that allowed information to be shared between branches. Whilst biologically persuasive, such non-homogeneous models are, however, highly parameterised and efforts have been made to seek more parsimonious representations. Yang & Roberts (1995) and Foster (2004) both considered models in which composition vectors are applied to groups of edges rather than to a single edge. Blanquart & Lartillot (2006) used a variation of this idea by assuming the compositional shifts occurred according to a Poisson process, independently of speciation events. In the context of nucleotide evolution, Galtier & Gouy (1998) reduced the number of parameters in the model of Yang & Roberts (1995) by using a model parameterised by a single G+C component, rather than three free parameters for the composition vector. But this inevitably came at the cost of a loss of information from the alignment. In a general setting that allowed different reversible or non-reversible rate matrices to be assigned to each edge of the tree, Jayaswal *et al.* (2011) devised a heuristic to reduce the number of rate matrices using the distances between them as a similarity criteria, and forcing the most similar rate matrices to be identical. However, given the speculative nature of the model search, the algorithm offered no assurance of identifying a global optimum.

1.4 Contribution

In spite of its importance, rooting major cellular radiations remains an under-investigated and challenging area of phylogenetics. The assumptions of standard phylogenetic models make them unable to infer rooted trees. Models which allow root inference are typically non-homogeneous, assigning different rate matrices to different parts of the tree. While being more realistic from a biological point of view, these models are substantially more highly parameterised than their homogeneous counterparts. This makes model-fitting challenging, often limiting inference to fixed unrooted trees or alignments on a small number of taxa.

We take a Bayesian approach to inference and focus on rooting using homogeneous models which require only one rate matrix. This approach has been explored previously by Huelsenbeck *et al.* (2002). Here we build on that work in a number of ways. First, we do not fix the unrooted topology and extend the inferential algorithm to allow inference of rooted trees. This allows us to present a more complete summary of the posterior over root positions and to demonstrate the sensitivity of the analysis to different topological priors. Additionally, Huelsenbeck *et al.* (2002) used a so-called non-informative prior on the rate matrix, with independent uniform distributions for each off-diagonal element. We incorporate prior structure and consider two hierarchical models which are centred on a standard reversible rate matrix but allow non-reversible perturbation of the individual elements. The two models differ in the structure of the perturbation. We also investigate non-stationary models in which the initial distribution at the root of the tree differs from the theoretical stationary distribution.

1.5 Overall structure

The thesis investigates rooting phylogenetic trees using model-based approaches. The analysis is performed in the Bayesian framework, aiming at inferring rooted phylogenetic trees from aligned molecular sequences. Chapter 2 contains the necessary background on Bayesian phylogenetic inference, as well as standard models of sequence substitution.

Chapter 3 describes two non-reversible substitution models. The chapter consists of the description of the models, their implementation through the MCMC algorithm and a simulation study. The simulations explore the performance of the models for different levels of non-reversibility in the data simulated under a random rooted tree. We also investigate the effect of different topologies and branch lengths on root inference, as well as the sensitivity of root inference to different topological priors. We show that as the level of non-reversibility in the data increases, root inference improves. As far as the branch lengths are concerned our results show that long branches can potentially mislead the rooting inference if the prior favours short branches.

Chapter 4 deals with non-stationary models. It first analyses a model which is non-stationary and reversible. Non-stationarity is achieved by introducing a composition at the root vertex which differs from the theoretical stationary composition. We then combine the idea of non-stationarity with the non-reversible models from Chapter 3, thus obtaining models which are non-reversible and non-stationary. Simulations are used to investigate the behaviour of the models for different levels of non-stationarity and non-reversibility in the data, as well as different topologies and different alignment lengths.

Chapter 5 focuses on applying the models to experimental data. First we apply our models to real biological data sets for which there is a robust biological opinion about the position of the root: the palaeopolyploid yeasts and the primates. We explore the composition of the nucleotides in experimental data and perform posterior predictive simulations. We show that while non-reversible models are able to extract some information about the root, modelling non-stationarity with just two composition vectors can be misleading for certain data sets. We then apply our models to an open question in biology: the root of the tree of life. Our results are in accord with the current biological opinion about the tree of life, whereby eukaryotes have originated within the Archaea, and the root is located either on the branch leading to the Bacteria, or within the Bacteria.

Chapter 6 summarises the thesis and outlines potential directions of further development.

Chapter 2

Background

2.1 Markov models of substitution

2.1.1 Markov process

Markov models of substitution aim to model the evolutionary process operating along each edge of a phylogenetic tree by approximating the processes of change from one nucleotide (or amino acid) of the ancestor to another one of the descendants over some period of time. Let us consider a single site of a DNA sequence. The nucleotide at this site can be thought of as a realisation of a random variable $X(t)$ indexed by time t that adopts values in a discrete finite space $\Omega = \{\text{A, G, C, T}\}$. The substitution process at the site is described by a continuous time Markov process, where the characters at the site are the states of the process. A Markov process is a stochastic process with the property that, given the current state, the future states do not depend on the past states. In other words, the probability of a nucleotide changing depends on its current value only and does not depend on its past values given this current value:

$$\begin{aligned}\Pr(X(t_n) = i_n | X(t_{n-1}) = i_{n-1},\ X(t_{n-2}) = i_{n-2}\ , \ldots,\ X(t_1) = i_1) \\ = \Pr(X(t_n) = i_n | X(t_{n-1}) = i_{n-1}),\end{aligned}$$

for any $t_n > t_{n-1} > t_{n-2} > \ \ldots \ > t_2 > t_1$.

The process can therefore be specified by a transition probability matrix $P(t) = \{p_{ij}(t)\}$ whose elements p_{ij} represent the probabilities of changing from one nucleotide to another during time period t:

$$p_{ij}(t) = \Pr(X(t) = j | X(0) = i).$$

Every row of the transition probability matrix sums to one. The transition probability

matrix is characterised by the Chapman-Kolmogorov equations:

$$p_{ij}(t_1 + t_2) = \sum_{k \in \Omega} p_{ik}(t_1) p_{kj}(t_2). \tag{2.1}$$

This means that the probability of changing from state i to state j is a sum of probabilities of changing from the state i to the intermediate state k, and then from the intermediate state k to the target state j (the sum is over all intermediate states $k \in \Omega$). If $P(t)$ is differentiable then

$$P(t) = P(0) + tQ + O(t^2)$$

is a Taylor expansion of $P(t)$ about $t = 0$, so that

$$Q = \lim_{t \to 0} \frac{P(t) - I}{t},$$

where $I = P(0)$. Thus $Q = dP/dt$ evaluated at $t = 0$. The matrix Q is called an *instantaneous rate matrix*. The off-diagonal elements of Q represent an instantaneous rate of change from one nucleotide to another during an infinitesimal period of time. It can be shown that the diagonal elements of Q are specified such that every row sums to zero:

$$\sum_j q_{ij} = \sum_j \lim_{t \to 0} \frac{p_{ij}(t) - \delta_{ij}}{t} = 0,$$

since

$$\sum_j p_{ij} = \sum_j \delta_{ij} = 1,$$

where

$$\delta_{ij} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases}$$

For instance, a rate matrix for DNA substitution can be represented as follows:

$$Q = (q_{ij}) = \begin{pmatrix} -(q_{12} + q_{13} + q_{14}) & q_{12} & q_{13} & q_{14} \\ q_{21} & -(q_{21} + q_{23} + q_{24}) & q_{23} & q_{24} \\ q_{31} & q_{32} & -(q_{31} + q_{32} + q_{34}) & q_{34} \\ q_{41} & q_{42} & q_{43} & -(q_{41} + q_{42} + q_{43}) \end{pmatrix},$$

where $q_{ij} \geqslant 0$ for all $i \neq j$. The relationships between P and Q at general time t are determined by forward and backward Kolmogorov equations which can be derived from (2.1):

$$P(t + h) = P(t)P(h) = P(t)\{I + hQ + O(h^2)\}, \text{ where } h \to 0.$$

It follows that

$$\frac{P(t+h)-P(t)}{h} = P(t)Q$$

and

$$\frac{P(t+h)-P(t)}{h} = QP(t).$$

Thus forward and backward Kolmogorov equations are given by

$$\frac{dP(t)}{dt} = P(t)Q$$

and

$$\frac{dP(t)}{dt} = QP(t)$$

with the solution

$$P(t) = I + tQ + \frac{1}{2!}t^2Q^2 + \frac{1}{3!}t^3Q^3 + \cdots = \exp(Qt).$$

The rate matrix Q can be written in a diagonal form:

$$Q = U \times \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \times U^{-1},$$

where the λ_i are eigenvalues of Q, and the columns of U are eigenvectors of Q. Then the transition probability matrix can be calculated using the diagonal form of Q:

$$P(t) = \exp(Qt) = U \times \mathrm{diag}(\mathrm{e}^{\lambda_1 t}, \ldots, e^{\lambda_n t}) \times U^{-1}.$$

2.1.2 Stationary distributions

The Markov process operating along each edge of the tree allows any state to change into any other state in finite time with positive probability, that is any nucleotide can be replaced by any other nucleotide. Such a process is called irreducible, and has a unique stationary distribution, i.e. the distribution of the nucleotides after a long time has elapsed:

$$\lim_{t\to\infty} p_{ij}(t) = \pi_j$$

for all i. In other words it is the distribution of the states after an infinitely large number of substitutions have occurred such that it is independent of the starting distribution of the states. The stationary distribution satisfies the equation

$$\boldsymbol{\pi} = \boldsymbol{\pi}P(t) \tag{2.2}$$

for all t. If $X(t_1)$ has distribution $\boldsymbol{\pi}$, then $X(t_1 + t_2)$ will have distribution $\boldsymbol{\pi}$ for all positive values of t_2. The stationary distribution can be derived from the rate matrix by differentiating equation (2.2) with respect to t: $\boldsymbol{\pi}Q = \mathbf{0}$.

A stationary distribution $\boldsymbol{\pi}$ is a row eigenvector of the rate matrix Q with eigenvalue 0. Equivalently, $\boldsymbol{\pi}$ is a row eigenvector of the transition matrix P with eigenvalue 1. A proof of the existence of the stationary distribution for 4-by-4 irreducible substitution models is given in Appendix A.

2.1.3 Substitution model on a tree

The transition probability matrix over an edge e of a phylogenetic tree is

$$P_e = \exp(\mu_e t_e Q_e)$$

where t_e is a time duration, μ_e is a rate of substitution events and Q_e is a normalised instantaneous rate matrix ($Q_e = Q/\rho_Q$, where $\rho_Q = -\sum_i q_{ii}\pi_i$ is an overall substitution rate). The edge length $\ell_e = \mu_e t_e$ represents an expected number of substitution events during time t_e.

2.1.4 Likelihood of a phylogenetic tree

Single branch

Let us consider the likelihood of a rooted phylogeny T with just a single branch e of length ℓ_e, where x is is the observed nucleotide of a single site of DNA sequence, and i is the unobserved ancestral state (the root) (Figure 2.1). Assuming the process is in the stationary distribution $\boldsymbol{\pi}$, the probability that the nucleotide at the root has value i is π_i. The likelihood of the tree is then $\pi(x|\tau) = \sum_i \pi_i \times p_{ix}(\ell_e)$.

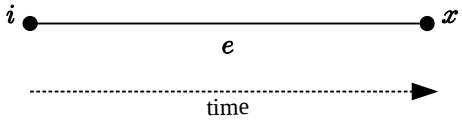


Figure 2.1: Phylogeny comprising a single branch e of lengths ℓ_e, where x is the observed nucleotide and i is the unobserved ancestral state (the root).

Many branches

Let us consider a single site of DNA sequence evolving along a rooted phylogeny τ with vertices V and edges E such that each edge is of the form $e = (v, w)$, where $v, w \in V$. Denote by $X(i)$ a nucleotide at vertex i which is only observed at the leaves. We assume that sequences change according to a Markov substitution model along each edge of the

tree, and the processes are the same on separate edges. Suppose the nucleotides at the internal vertices are known, in this case the probability of the data is a product of the transition probabilities associated with every edge of the tree multiplied by the probability associated with the root vertex:

$$\pi(x|X,\tau) = \pi_{X(root)} \prod_{\text{edges } e=(v,w)} p_{X(v),X(w)}(\ell_e),$$

where ℓ_e is the length of the edge e. However, the nucleotides at the internal vertices are unobserved. The likelihood therefore is obtained by averaging over all possible unobserved nucleotide values at the internal vertices and at the root vertex:

$$\pi(x|\tau) = \sum_{X} \pi_{X(root)} \prod_{\text{edges } e=(v,w)} p_{X(v),X(w)}(\ell_e).$$

The sum is taken over all functions $X(i)$ from the vertices to Ω such that $X(i)$ matches data x_i for leaf vertices. For example, the likelihood of the tree depicted in Figure 2.2 is:

$$\pi(x|\tau) = \sum_{X_{(root)}} \sum_{X(5)} \sum_{X(6)} \pi_{X(root)} p_{X(5),T}\; p_{X(5),C}\; p_{X(6),A}\; p_{X(6),C}\; p_{X(0),X(5)}\; p_{X(0),X(6)}. \tag{2.3}$$

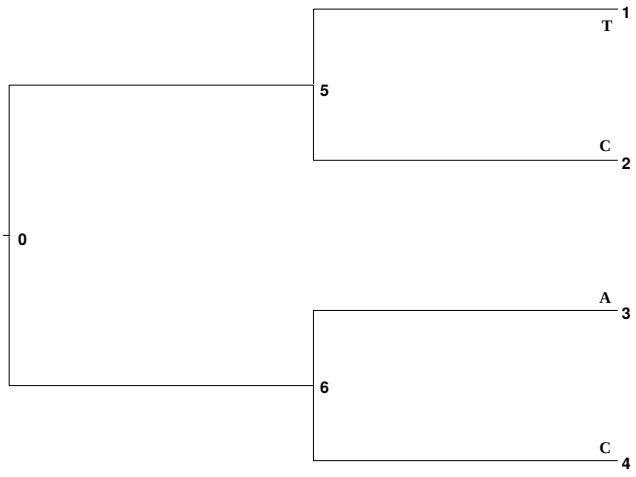


Figure 2.2: Four-taxon tree for evaluating the likelihood according to equation (2.3).

Since the evaluation of the likelihood grows like $|\Omega|^N$ for N taxa, it is convenient to use the Felsenstein pruning algorithm (Felsenstein, 1973, 1981) to save computational time. The algorithm works by calculating the conditional probabilities at each node of the tree recursively from the tips of the tree towards the root, as illustrated in the following example. Suppose we are interested in evaluating the likelihood of observing the nucleotides

T, C, A, C at the leaves 1, 2, 3, 4 of the tree depicted in Figure 2.2 and we are given the following transition probability matrix:

$$P(\ell) = \begin{pmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.25 & 0.25 & 0.2 & 0.3 \\ 0.22 & 0.26 & 0.28 & 0.24 \\ 0.32 & 0.13 & 0.27 & 0.28 \end{pmatrix}.$$

Suppose for simplicity that all the edges of the tree have length ℓ. The vector of the conditional probabilities $\mathbf{V}_5$ for the node 5 is calculated as follows:

$$\begin{aligned} V_5(A) &= p_{AT} \times p_{AC} = 0.3 \times 0.4 = 0.012, \\ V_5(G) &= p_{GT} \times p_{GC} = 0.2 \times 0.3 = 0.06, \\ V_5(C) &= p_{CT} \times p_{CC} = 0.24 \times 0.28 = 0.0672, \\ V_5(T) &= p_{TT} \times p_{TC} = 0.28 \times 0.27 = 0.0756. \end{aligned}$$

The vector $\mathbf{V}_6 = (0.03,\ 0.05,\ 0.0616,\ 0.0864)$ is calculated similarly. The conditional probabilities at the node 0 are calculated using the vectors $\mathbf{V}_5$ and $\mathbf{V}_6$ (Figure 2.3). For example,

$$\begin{aligned} V_0(A) &= (p_{AA} \times V_5(A) + p_{AG} \times V_5(G) + p_{AC} \times V_5(C) + p_{AT} \times V_5(T)) \\ &\quad \times (p_{AA} \times V_6(A) + p_{AG} \times V_6(G) + p_{AC} \times V_6(C) + p_{AT} \times V_6(T)) \\ &= (0.1 \times 0.012 + 0.2 \times 0.06 + 0.3 \times 0.0672 + 0.4 \times 0.0756) \\ &\quad \times (0.1 \times 0.03 + 0.2 \times 0.05 + 0.3 \times 0.0616 + 0.4 \times 0.0864) \\ &= 0.004200144. \end{aligned}$$

After calculating the conditional probabilities of all the nodes on the tree, the probability of the data is computed by

$$\pi(x|\tau) = \sum_{X_{(root)}} \pi_{X(root)} V_0(X).$$

Many sites

Let us consider an alignment comprising n sites. Since we assume that the sites of the tree evolved independently of each other, the likelihood can be expressed as a product of the likelihoods of n individual sites of the alignment:

$$\pi(D|\tau) = \prod_{i=1}^{n} \pi(D_i|\tau),$$

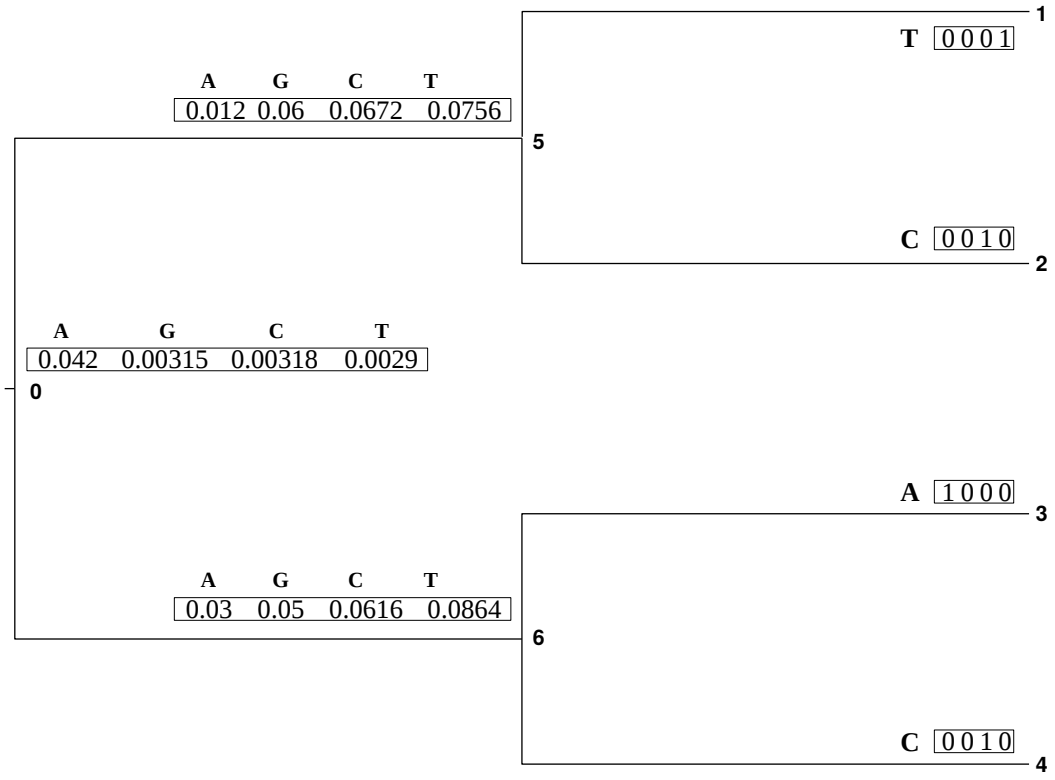


Figure 2.3: Example of calculating the conditional probabilities at nodes according to the Felsenstein pruning algorithm.

where $\pi(D_i|\tau)$ is the likelihood of an individual site.

2.1.5 Homogeneity, stationarity, reversibility

One of the common assumptions of phylogenetics is that the evolutionary process at each site is homogeneous, stationary and reversible. Homogeneity implies that a single instantaneous rate matrix Q applies to the whole tree. Stationarity implies that the probability of each nucleotide does not change over time and the probabilities of transitioning between nucleotides over some time interval depend only on the size of that interval and not on its position in time. If the model is homogeneous and stationary then it might or might not be reversible. Reversibility allows the rate matrix to be represented in the form $Q = S\Pi$, where S is a symmetric matrix of the exchangeability parameters $S = (\rho_{ij})$, and $\Pi = \text{diag}(\boldsymbol{\pi})$ is a diagonal matrix containing the elements of $\boldsymbol{\pi}$:

$$Q = (q_{ij}) = \begin{pmatrix} \star & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & \star & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & \star & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & \star \end{pmatrix} \times \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_G & 0 & 0 \\ 0 & 0 & \pi_C & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix}.$$

The diagonal elements of S are specified such that every row sums to zero. Reversibility is described by the detailed balance equation:

$$\pi_i p_{ij}(\ell) = \pi_j p_{ji}(\ell). \tag{2.4}$$

Reversibility implies that the probability of sampling nucleotide i from the stationary distribution and going to nucleotide j is equal to that of sampling nucleotide j from the stationary distribution and going to nucleotide i (Felsenstein, 1981).

Reversibility leads to an important implication as far as the likelihood function is concerned: changing the root position does not change the likelihood of the tree. We will use a simple example to demonstrate how the likelihood under the reversibility condition does not depend on the root position. Consider a tree with just two taxa where G and A are the nucleotides at the leaves. Suppose that there are two alternative rooting positions for this tree: rooting at the vertex X and rooting at the vertex Y (Figure 2.4). We will

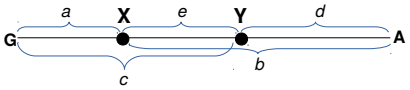


Figure 2.4: A tree with two taxa, where G and A are the nucleotides at the leaves. Vertices X and Y represent two possible root positions.

show that under the reversibility condition the likelihood of the tree is the same regardless of whether it is rooted at the vertex X or Y.

The probability of observing this tree with the root at the vertex X is

$$L_1 = \sum_{X \in \Omega} \pi_X p_{XG}(a) p_{XA}(b),$$

while the probability of observing this tree with the root at the vertex Y is

$$L_2 = \sum_{Y \in \Omega} \pi_Y p_{YG}(c) p_{YA}(d).$$

According to the Chapman-Kolmogorov equation (2.1) the term $p_{YG}(c)$ can be re-arranged as $\sum_{X \in \Omega} p_{YX}(e) p_{XG}(a)$. According to the detailed balance equation (2.4),

$$p_{YX}(e) = \frac{1}{\pi_Y} p_{XY}(e) \pi_X,$$

such that

$$p_{YG}(c) = \sum_{X \in \Omega} \frac{1}{\pi_Y} p_{XY}(e) \pi_X p_{XG}(a).$$

Now we substitute the term $p_{YG}(c)$ into L_2:

$$L_2 = \sum_{Y \in \Omega} \pi_Y \sum_{X \in \Omega} \frac{1}{\pi_Y} p_{XY}(e) \pi_X p_{XG}(a) p_{YA}(d).$$

According to the Chapman-Kolmogorov equation, the terms coloured in red correspond to $p_{XA}(b)$, and the terms π_Y and $1/\pi_Y$ are cancelled out, so

$$L_2 = \sum_{X \in \Omega} \pi_X p_{XG}(a) p_{XA}(b) = L_1.$$

Thus the likelihood of the tree is the same regardless of whether it is rooted at the vertex X or Y. This means that the reversibility condition allows us to ignore the direction of evolution since changing the placement of the root does not change the likelihood.

While the homogeneity, stationarity and reversibility assumptions make statistical models simpler, they have no biological justification, and are applied for computational convenience only. Evidence of non-stationarity and non-reversibility has indeed been found in biological data sets (Squartini & Arndt, 2008).

2.1.6 Molecular clock

We have shown in the previous section that imposing the reversibility constraint does not allow inference of rooted trees. However this is only true if the molecular clock is not assumed. Recall from Chapter 1 that the molecular clock assumption implies that the rate of substitution events μ is constant over time. Under the clock the branch lengths on the tree are proportional to time, therefore every leaf is equidistant from the root. The constraint of equal distance of the tips from the root makes it possible to identify the root, since re-rooting will violate the constraint, as illustrated in Figure 2.5.

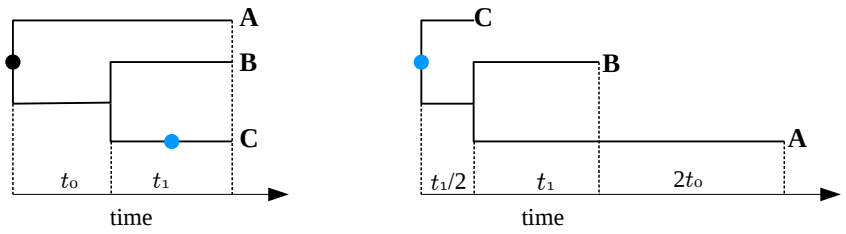


Figure 2.5: An illustration of identifying the root using a molecular clock. The tree on the left hand side is constructed according to the molecular clock assumption (the leaves are equidistant from the root mapped with a black circle). Re-rooting the tree on the middle of the branch leading to the leaf C (blue circle) creates a tree which violates the molecular clock assumption (the tree on the right hand side, where the leaves are not equidistant from the root).

2.1.7 Majority rule consensus tree

One of the common ways to summarise the information contained in a set of trees is a *majority rule consensus tree*. A rooted majority rule consensus tree is constructed in a way that it contains clades which appear in at least 50% of the analysed set of trees. For example, suppose we are interested in constructing a rooted majority rule consensus tree from the set of four rooted trees: Tree 1, Tree 2, Tree 3 and Tree 4 in Figure 2.6. Clades (A, B) and (C, D) appear on two trees (Tree 1 and Tree 2) therefore they are included in the consensus tree with associated probability 0.5. The clade (A, B, C, D) appears on the Trees 3 and 4, so it appears on the consensus tree with associated probability 0.5. It is worth noting that the consensus tree does not have to represent one of the analysed trees. Indeed, in our example the consensus tree is different from the four trees in the set.

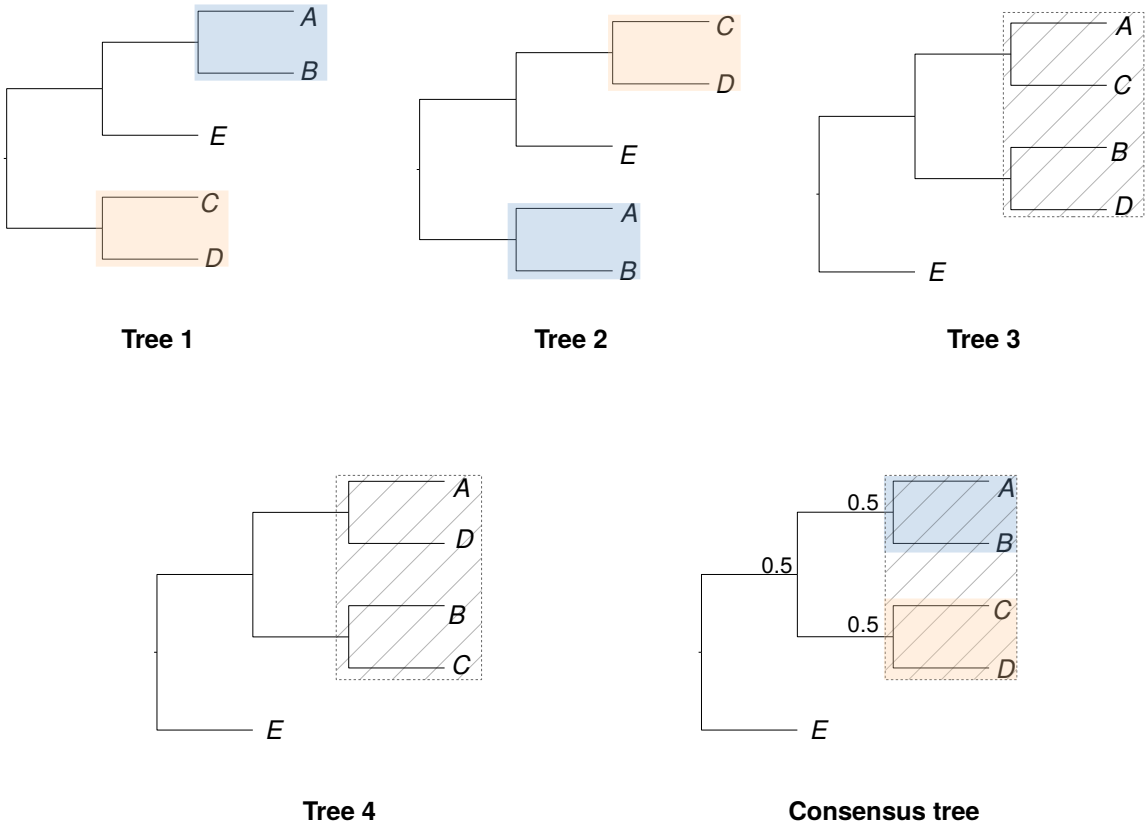


Figure 2.6: Rooted majority rule consensus tree, constructed from Tree 1, Tree 2, Tree 3 and Tree 4. The clades (A, B) and (C, D) appear on Trees 1 and 2, the clade (A, B, C, D) appears on Trees 3 and 4. The three clades therefore appear on the consensus tree, each one with associated probability 0.5. Note that the consensus tree is different from all the four trees.

2.2 Models of nucleotide substitution

2.2.1 The JC69 model

The JC69 model (Jukes & Cantor, 1969) is the simplest substitution model which assumes equal rates of substitution between any two nucleotides, so that the rate matrix is given by

$$Q = (q_{ij}) = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$

for some constant λ. This model has symmetrical substitution rates ($q_{ij} = q_{ji}$), meaning that the stationary distribution is 1/4 for every nucleotide.

2.2.2 The K80 model

The K80 model (Kimura, 1980) allows different substitution rates for transitions (substitutions between two purines or between two pyrimidines : $A \leftrightarrow G$ or $C \leftrightarrow T$) and transversions (substitutions between purines and pyrimidines: $A, G \leftrightarrow C, T$). Let the substitution rates be α for transitions and β for transversions. The rate matrix is represented as follows:

$$Q = (q_{ij}) = \begin{pmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \alpha & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{pmatrix}.$$

This model is more realistic than the JC69 because in real data transitions often occur at higher rates than transversions (Brown *et al.*, 1982; Gojobori *et al.*, 1982; Curtis & Clegg, 1984). However, the stationary distribution in this model is again 1/4 for every nucleotide.

2.2.3 The TN93 model

The TN93 model (Tamura & Nei, 1993) relaxes the assumption of equal stationary probabilities for the four nucleotides. Representing the stationary distribution by the vector $\boldsymbol{\pi} = (\pi_A,\ \pi_G,\ \pi_C,\ \pi_T)$, the rate matrix is expressed as

$$Q = (q_{ij}) = \begin{pmatrix} -(\alpha_1\pi_G + \beta\pi_Y) & \alpha_1\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha_1\pi_A & -(\alpha_1\pi_A + \beta\pi_Y) & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & -(\alpha_2\pi_T + \beta\pi_R) & \alpha_2\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha_2\pi_C & -(\alpha_2\pi_C + \beta\pi_R) \end{pmatrix},$$

where α_1, α_2, and β represent the rates of transitional changes between purines and between pyrimidines and of transversional change, respectively. It is worth noting that even though the rate matrix is not symmetric, the model is still reversible since it satisfies the detailed balance equation (2.4).

2.2.4 The HKY85 model

The HKY85 model (Hasegawa *et al.*, 1985) can be considered a special case of the TN93 model obtained by setting $\alpha_1 = \alpha_2 = \alpha$. It can be also parameterised by fixing β at the value 1 and defining a transition-transversion rate ratio by $\kappa = \alpha/\beta$. The rate matrix is represented as follows:

$$Q = (q_{ij}) = \begin{pmatrix} -(\pi_G + \kappa\pi_C + \pi_T) & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & -(\kappa\pi_A + \pi_C + \pi_T) & \pi_C & \pi_T \\ \pi_A & \pi_G & -(\pi_A + \pi_G + \kappa\pi_T) & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & -(\pi_A + \pi_G + \kappa\pi_C) \end{pmatrix}.$$

2.2.5 The GTR model

The general time reversible model (Tavaré, 1986) has different instantaneous rates of substitution between each of the six nucleotide pairs, while the reversibility condition still holds. In fact, all other models are special cases of the GTR model which are achieved by assuming equality amongst some of the parameters. Representing the rate parameters, sometimes called exchangeability parameters, by vector $\boldsymbol{\rho} = (\rho_{ij})$, $i = 1, 2, 3$, $j = i + 1, \ldots, 4$ the rate matrix can be represented as follows:

$$Q = (q_{ij}) = \begin{pmatrix} -(\rho_{12}\pi_G + \rho_{13}\pi_C + \rho_{14}\pi_T) & \rho_{12}\pi_G & \rho_{13}\pi_C & \rho_{14}\pi_T \\ \rho_{12}\pi_A & -(\rho_{12}\pi_A + \rho_{23}\pi_C + \rho_{24}\pi_T) & \rho_{23}\pi_C & \rho_{24}\pi_T \\ \rho_{13}\pi_A & \rho_{23}\pi_G & -(\rho_{13}\pi_A + \rho_{23}\pi_G + \rho_{34}\pi_T) & \rho_{34}\pi_T \\ \rho_{14}\pi_A & \rho_{24}\pi_G & \rho_{34}\pi_C & -(\rho_{14}\pi_A + \rho_{24}\pi_G + \rho_{34}\pi_C) \end{pmatrix}.$$

2.3 Models of amino acid substitution

It is straightforward to apply the continuous-time Markov process to describe substitutions between amino acids. The states of the process now comprise 20 amino acids: $\Omega = \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$. Amino acid models are generally empirical models, that is, various parameters are fixed at values based on analyses of large quantities of sequence data. The models are constructed by estimating

relative rates between amino acids under the GTR model. Thus, the rate matrix of the substitution model for amino acids is expressed by $Q = S\Pi$, where $S = (s_{ij})$ is a 20-by-20 matrix of the amino acid exchangeabilities, and $\Pi = \text{diag}(\pi_1,\ \pi_2,\ \ldots,\ \pi_{20})$ is the equilibrium frequency of the amino acids. Typically Π is regarded as unknown and S is fixed. A number of amino acid exchangeabilities matrices have been proposed. The first empirical amino acid substitution matrices were constructed by averaging the number of amino acid changes in closely related sequence pairs (Dayhoff *et al.*, 1989; Jones *et al.*, 1992). The WAG matrix is based on the phylogeny of sequences rather than on the counting method (Whelan & Goldman, 2001). The LG matrix takes into account the variability of the evolutionary rates across sites (Le & Gascuel, 2008).

2.4 Models with a reduced alphabet (Dayhoff re-coding)

These models deal with groups of amino acids rather than with individual amino acids. This approach reduces the alphabet from the number of amino acids (20) to a number of groups. There are different approaches of clustering amino acids into groups. One of the approaches called Dayhoff re-coding is based on looking for groups of chemically related amino acids that commonly replace one another. Using this approach, the amino acids are clustered into six groups: *AGPST*, *DENQ*, *HKR*, *ILMV*, *FWY* and *C*. Each group of amino acids is treated as the same single character state, and hence this method has an effect on homogenising the amino acid composition between sequences (Hrdy *et al.*, 2004). The GTR model is often used.

Re-coding of the amino acids into groups of amino acids helps to avoid the *saturation* problem. Saturation occurs when multiple substitutions obscure the phylogenetic signal such that it is no longer possible to accurately estimate sequence divergence. Since the changes within the groups are much more common than the changes between the groups, treating a group as a character can be beneficial for avoiding saturation in substitution events (Embley *et al.*, 2003; Susko & Roger, 2007).

2.5 Across-site heterogeneity

The rates of substitution can vary among sites due to their different roles in the structure and function of the gene. Rates of nucleotide substitution at different sites are highly variable in most genes because of the existence of variable and conserved regions in the gene (Yang & Roberts, 1995). For example, sites which perform fundamental roles in life and exist in all organisms are more likely to be conserved, while other sites may accumulate very many changes. The rates among sites can be accommodated by assuming that each site i has its own rate r_i represented by a random variable drawn from a statistical distribution

(a gamma distribution is often used) (Yang, 1993). The rate matrix for a site with rate r_i is then r_iQ and we assume that r_i is equal to one on average. Often it is assumed that $r_i|\alpha \sim \text{Ga}(\alpha, \alpha)$. Thus, the distribution has a mean of 1 and the variance of $1/\alpha$, and manipulating α allows manipulation of the shape of the distribution. For instance, if $\alpha > 1$ then the distribution is bell-shaped, meaning that most of the sites have rates around 1 with few sites having very low or very high rates. If $\alpha \leq 1$ then the distribution has a skewed L-shape, meaning that most sites have very low rates of substitution and only a few sites have high rates (Yang, 2006) (Figure 2.7). The parameter α is generally treated as an unknown parameter about which inference is sought.

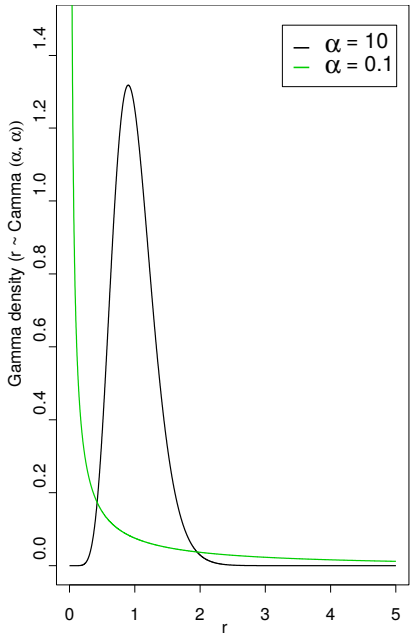


Figure 2.7: Shapes of $\text{Ga}(\alpha, \alpha)$ for two different values of α ($\alpha > 1$ and $\alpha < 1$). For $\alpha = 10$ (black line) the distribution is concentrated around 1 meaning that very few sites have low or high rates. For $\alpha = 0.1$ (green line) most of the sites have very low rates.

Since this model is expensive computationally, a discrete-gamma model has been suggested whereby several equal-probability categories are used to approximate the continuous gamma distribution, with the mean of each category representing all rates in that category (Yang, 1994). Suppose there are n categories with rates $r_1, \ldots, r_n$ and the probability of each category is $p_k = 1/n, \quad k = 1, \ldots, n$. Then the likelihood of the data at a site i is:

$$P(D_i|\boldsymbol{\theta}) = \sum_{k=1}^{n} p_k \times P(D_i|r_k, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ represents the parameters of the substitution model, and r_k is the rate of the k-th category. By analysing goodness of fit of models with different numbers of categories,

Yang (1994) suggested using $n = 4$. Even though there is no biological reason to use a gamma distribution for among site rate variation, it allows a wide range of rate shapes with only a single parameter.

There are other approaches of modelling heterogeneity across sites which allow other model features to vary across sites. Lartillot & Philippe (2004) considered a mixed model of K distinct classes, each class is characterised by its own substitution matrix Q_K with fixed exchangeability parameters $\boldsymbol{\rho}$, such that the mixture is defined on the space of stationary distributions $\boldsymbol{\pi}$. Working in a Bayesian framework, the model utilises a Dirichlet process prior for the stationary probabilities for each class, with the number of classes being a free parameter in the model. A stochastic allocation vector gives a probability of assigning a site to each possible class. Conditioning on a fixed unrooted topology, another approach (Jayaswal *et al.*, 2014) considers a fixed assignments model, in which prespecified groups of sites are assigned their own rate matrix. Given a particular number of sites, each allocation is considered to be a different model. Performing a bottom-up search by increasing the number of groups, the optimal model is specified as the best fitting model. Working in a maximum likelihood framework, Jayaswal *et al.* (2007) considered a mixture model of variable and invariant (constant over time) sites, with the variable sites evolving according to the same Markov process. Thus, heterogeneity across variable sites was not considered.

2.6 Bayesian inference

2.6.1 Overview

According to Bayesian statistics, inference about an event is made using prior belief as well as data, that is knowledge or experience about how likely the event is to happen. Thus, the parameters are considered to be random variables with statistical distributions rather than unknown fixed constants as in the frequentist approach. Before the analysis, the parameters are assigned prior distributions, which are combined with the likelihood of the data to generate a posterior distribution. The posterior distribution represents the uncertainty about the parameters after observing the data and is calculated from the prior distribution modified by the likelihood of the data. All inferences concerning the parameters are then based on the posterior distribution of the parameters. According to Bayes' theorem, the posterior probability is:

$$\pi(\theta|D) = \frac{\pi(\theta)f(D|\theta)}{f(D)}, \tag{2.5}$$

where $\pi(\theta)$ is the probability density function of the parameter θ, $f(D|\theta)$ is the likelihood function of the data given the value of θ, $\pi(\theta|D)$ is the posterior density of θ given the data, and $f(D) = \int_\theta f(D|\theta)\pi(\theta)\mathrm{d}\theta$ is the marginal probability of the data, which can be

thought of as a normalising constant.

Let us consider the following example. Suppose, an experiment has been done to identify cancerous cells. A stain has been developed which adheres only to the cancerous cells. We are interested in modelling the random variable X which represents the number of cancerous cells in a sample. We can use the binomial $\text{Bin}(n, p)$ distribution to describe the data, where n is the number of cells, and p is the probability of a cell being cancerous. According to our prior belief the probability of a cell being cancerous is quite small, so we assign it a beta distribution with the mode 0.25: $p \sim \text{Beta}(2, 4)$. The prior probability density function of p is

$$\pi(p) \propto p(1-p)^3.$$

Suppose we collect a sample in which 2 out of the 20 cells are stained after the experiment. The likelihood of the data is $f(X|p) \propto p^x(1-p)^{20-x}$. The posterior distribution of p is then

$$\begin{aligned}\pi(p|X) &\propto \pi(p) \times f(X|p) \\ &\propto p(1-p)^3 \times p^x(1-p)^{20-x} \\ &\propto p^{x+1}(1-p)^{23-x},\end{aligned}$$

which is Beta(4, 22). The beta prior is called a conjugate prior for the binomial likelihood because the prior and the posterior are from the same family of distributions. This example illustrates that in Bayesian statistics the role of the data is to update the prior distribution of the parameters. The prior density, the likelihood and the posterior density are illustrated in Figure 2.8.

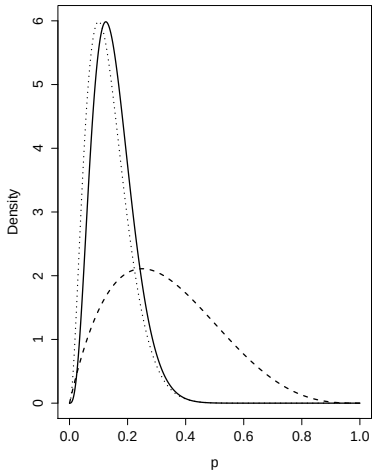
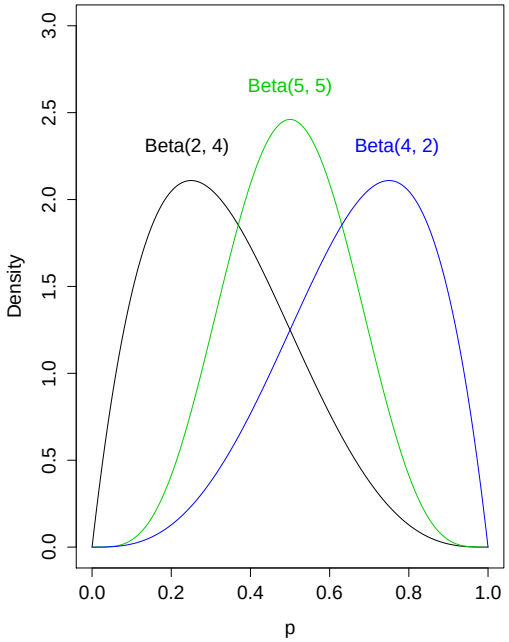
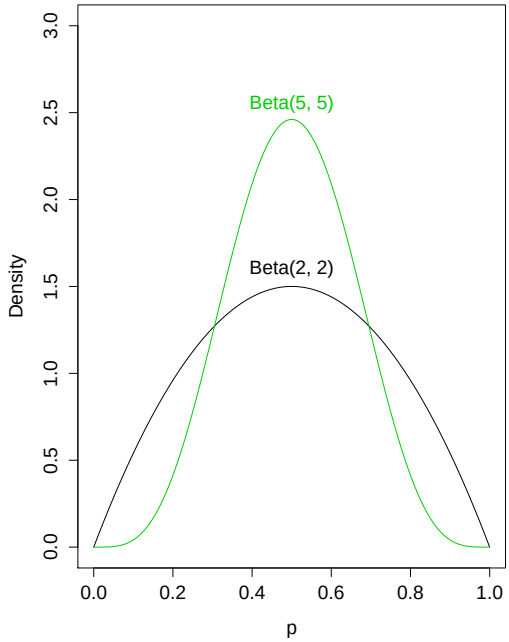


Figure 2.8: Prior density (dashes), likelihood (dots) and posterior density (solid) in the example of modelling the number of cancerous cells.

The choice of the prior distribution is an important issue in Bayesian statistics. It is convenient to select the prior on the basis of the distributional shape. For instance, in our example $p \sim \text{Beta}(5, 5)$ represents a prior belief that the probability of a cell being cancerous is the same as the probability of a cell being non-cancerous ($\text{E}(p) = 0.5$), while $p \sim \text{Beta}(4, 2)$ represents higher prior belief of a cell being cancerous ($\text{E}(p) = 2/3$) (Figure 2.9a). Equivalently, the prior uncertainty can be expressed by the the distributional shape. For instance, $p \sim \text{Beta}(2, 2)$ represents higher prior uncertainty than $p \sim \text{Beta}(5, 5)$, because the variance of Beta(2, 2) is bigger that the variance of Beta(5, 5) (0.05 and 0.023 respectively), while both distributions have the same mean 0.5 (Figure 2.9b).



(a) Beta distribution with different parameters: Beta(2,4), Beta (5,5) and Beta (4,2) with means 1/3, 1/2 and 2/3 respectively.



(b) Beta distribution with the same mean and different variances: the mean of Beta(5,5) and Beta(2,2) is 0.5, while the variances are 0.023 and 0.05 respectively.

Figure 2.9: Different shapes of beta distribution. Panel (a) shows three beta distributions with different means. Panel (b) shows two beta distributions with the same mean but different variances.

Sometimes we are interested in inferring only certain parameters. The parameters we are not interested in are called *nuisance parameters* and are dealt with through integration. Suppose $\boldsymbol{\theta} = (\mu,\ \sigma)$ are the parameters, where μ is the parameter of interest and σ is a nuisance parameter. The marginal posterior probability of μ is achieved by integrating out σ as follows:

$$\pi(\mu|D) = \frac{\pi(\mu)f(D|\mu)}{\int_\mu \pi(\mu)f(D|\mu)\mathrm{d}\mu} = \frac{\int_\sigma \pi(\mu,\sigma)f(D|\mu,\sigma)\mathrm{d}\sigma}{\int_\sigma \int_\mu \pi(\mu,\sigma)f(D|\mu,\sigma)\mathrm{d}\mu\mathrm{d}\sigma},$$

where $\pi(\mu)$ and $\pi(\mu,\sigma)$ are probability density functions of the parameters, $f(D|\mu)$ and $f(D|\mu,\sigma)$ are likelihood functions of the data given the values of the parameters.

2.6.2 Markov chain Monte Carlo

One of the challenges of Bayesian statistics is calculating the marginal likelihood of the data $f(D)$ (the denominator of equation (2.5)), which is problematic when no conjugate priors are available, or when the integration involves a large number of unknowns. In these cases the marginal likelihood is analytically intractable. For instance, in order to calculate the posterior probability of a phylogenetic tree the integration has to be performed over all parameters of the substitution model and branch lengths for every tree topology. Non conjugate Bayesian inference for problems with a large number of parameters was found to be impractical until the past two decades, when it has gained popularity due to development of advanced computational methods, especially Markov Chain Monte Carlo algorithms (MCMC algorithms) (Gilks *et al.*, 1996). Monte Carlo algorithms are computational algorithms for sampling from probability distributions. The idea behind MCMC is to construct a Markov chain, whose stationary distribution is the target posterior distribution, and then generate dependent samples from this distribution by sampling realisations from the Markov chain. Crucially, the normalising constant does not need to be calculated.

2.6.3 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm (Hastings, 1970) is one of the MCMC algorithms for sampling from the target density. Suppose we are interested in sampling from the posterior density $\pi(\theta|D)$. The algorithm utilises some proposal density $q(\theta'|\theta)$ called a transition kernel which is used to propose a new realisation θ' given the current realisation θ. The steps of the Metropolis-Hastings algorithm are as follows (Chib & Greenberg, 1995):
1. Initialise the iteration counter $i = 1$ and initialise the current state with some value θ_1 from the support of $\pi(\theta|D)$.
2. Generate a proposal value θ' using the transition kernel $q(\theta'|\theta_i)$.

3. Evaluate the acceptance probability α of the proposed move:

$$\begin{aligned}\alpha(\theta_i, \theta') &= \min\left\{1, \frac{\pi(\theta'|D)}{\pi(\theta_i|D)} \times \frac{q(\theta_i|\theta')}{q(\theta'|\theta_i)}\right\} \\ &= \min\left\{1, \frac{\pi(\theta')\pi(D|\theta')/f(D)}{\pi(\theta_i)\pi(D|\theta_i)/f(D)} \times \frac{q(\theta_i|\theta')}{q(\theta'|\theta_i)}\right\}.\end{aligned}$$

After cancelling the marginal likelihood of the data $f(D)$ in the numerator and the denominator, the acceptance probability simplifies as follows:

$$\alpha(\theta_i, \theta') = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta_i)} \times \frac{\pi(D|\theta')}{\pi(D|\theta_i)} \times \frac{q(\theta_i|\theta')}{q(\theta'|\theta_i)}\right\}.$$

4. Accept or reject the proposal value θ' based on the acceptance probability α. This step consists of generating a random number $u \sim \mathrm{U}(0, 1)$. If $u < \alpha(\theta_i, \theta')$ then θ' is accepted, otherwise it is rejected.
5. Set the new value θ_{i+1} of the chain: if the proposal is accepted, then $\theta_{i+1} = \theta'$. Otherwise $\theta_{i+1} = \theta_i$.
6. Go to step 2.

The generated samples are dependent. The posterior mean $\mathrm{E}(\theta|D)$ of a quantity θ may be approximated by the sample mean $\bar{\theta}$ of our sampled values of θ. As with any Monte Carlo method, the accuracy of the approximation is limited by the sampling variation inherent in taking random samples. This is measured by the Monte Carlo variance of $\bar{\theta}$. When the sampled values from successive iterations are positively autocorrelated, the Monte Carlo variance of $\bar{\theta}$ from n iterations is larger than it would be given a sample of n independent draws from the posterior distribution.

It is possible to reduce the autocorrelation across iterations by thinning the chain, that is by retaining only the sampled values from iterations $m, 2m, 3m, \ldots$ where m is an integer, $m > 1$. Thinning gives a sample of n/m values and an increased Monte Carlo variance but, when there is positive autocorrelation, the increase can be small. In cases where time-consuming computations are done on each sampled value after it is collected, it may be more computationally efficient to increase n and then thin using $m > 1$ before executing these post-sample computations. An example of this is the computation of posterior predictive means, as will be described in Section 5.1.3. Thinning can also be useful for assessing convergence and when storage space is a problem. These issues are discussed by Geyer (1992). The autocorrelation between samples can be monitored using the autocorrelation function (ACF) plot (Geyer, 2011) (Figure 2.10).

The transition kernel can be symmetric, that is $q(\theta'|\theta_i) = q(\theta_i|\theta') \ \forall \ \theta', \theta_i$ (Metropolis *et al.*, 1953). In this case the acceptance probability simplifies to the ratio of the prior

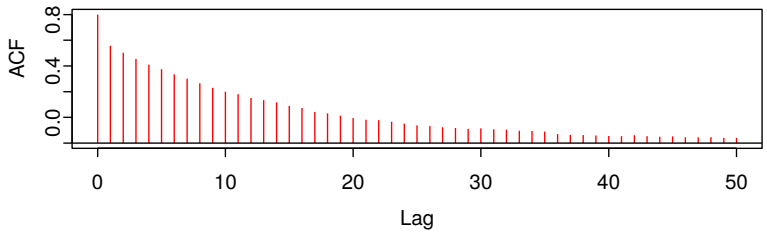


Figure 2.10: ACF plot showing the autocorrelation between the samples as a function of the iteration lag between them. The plot is displaying a decaying pattern of auto-correlation.

densities times the likelihood ratio, that is

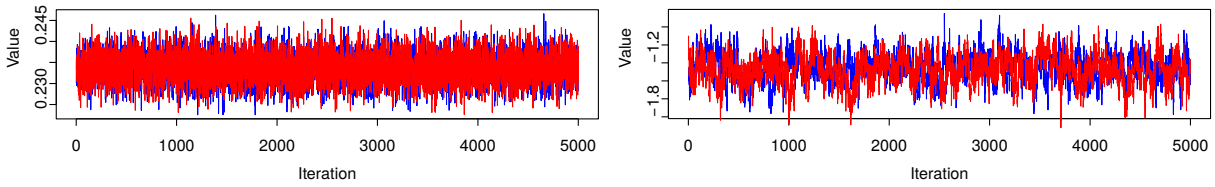
$$\alpha(\theta_i, \theta') = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta_i)} \times \frac{\pi(D|\theta')}{\pi(D|\theta_i)}\right\}.$$

The proposed value can be defined as a random variable from a distribution centred on the current state of the process. In this case the variance of the proposal distribution is controlling the size of the innovation. Large values of the variance are likely to cause most of the proposal values to be rejected, that is the process will remain at the same state for a long time, causing high autocorrelation. On the other hand, if the variance is too small, most of the proposal values will be accepted and the new states will be very close to the current state, leading again to the high autocorrelation. The variance can be tuned to achieve the desirable acceptance rate (Gilks *et al.*, 1996; Yang, 2006). It has been found experimentally that proposals leading to an acceptance rate around 30% minimise autocorrelation (Roberts *et al.*, 1997). Since the chain is initialised with some random value, it takes a certain amount of iterations until it reaches the stationary distribution. These iterations are called the burn-in period and they are discarded to ensure that the samples are drawn from the stationary distribution of the process (Brooks, 1998; Geyer, 2011).

Convergence diagnostics

The biggest concerns of an MCMC algorithm are *convergence* and *mixing*. "Convergence" means the ability of the chain to reach its stationary distribution (Yang, 2006). Often algorithms suffer from slow convergence, that is it takes a long time to reach stationarity. "Mixing" refers to how quickly the sampler explores the support of π. Poor mixing is indicated by high autocorrelation over iterations and inefficiency in exploring the parameter space. For instance, in the case of a multi-modal target distribution, the sampler might become stuck at one of the local modes. Thus, it is advisable to run multiple chains from different starting points and to make sure that they all converge to the same distribution (Gelman & Rubin, 1992). Though a variety of diagnostic tools has been proposed (Cowles

& Carlin, 1996) which utilise numerical methods based on either a single chain (Geweke, 1992) or multiple chains (Gelman & Rubin, 1992), one of the common methods is a visual examination of the plots of the parameters (Gelfand & Smith, 1990). For example, MCMC output can be diagnosed by trace plots in which the parameters are plotted against the iteration number (Figure 2.11). Stochastic nature of the MCMC algorithms imply that it



(a) (b)

Figure 2.11: Trace plots of a parameter from two different chains shown in red and blue colours. (a) Good mixing of the chains, represented by frequent moves around the support of the target distribution. (b) Poor mixing of the chains, suggested by a high autocorrelation.

is impossible to tell with certainty that convergence has been achieved. However, if the trace plots for any parameter sampled from two chains with different starting points fail to overlap, this is an indication of a lack of convergence. Therefore, when the process has many parameters it is important to monitor all of them. If even one of n parameters does not appear to have converged, the chain has not converged. After performing convergence diagnostics, the MCMC output can be summarised with respect to the parameters of interest.

2.6.4 Bayesian phylogenetics

In Bayesian phylogenetics the parameters are the substitution model parameters and the phylogenetic tree (including branch lengths), while the data are aligned sequences. The aim of Bayesian phylogenetics is to calculate the posterior probability of the phylogenetic tree, branch lengths and the parameters of the substitution model, that is probability of the tree and the parameters of the substitution model given the sequence data.

According to Bayes' theorem, the posterior density of the tree and the parameters of the substitution model is

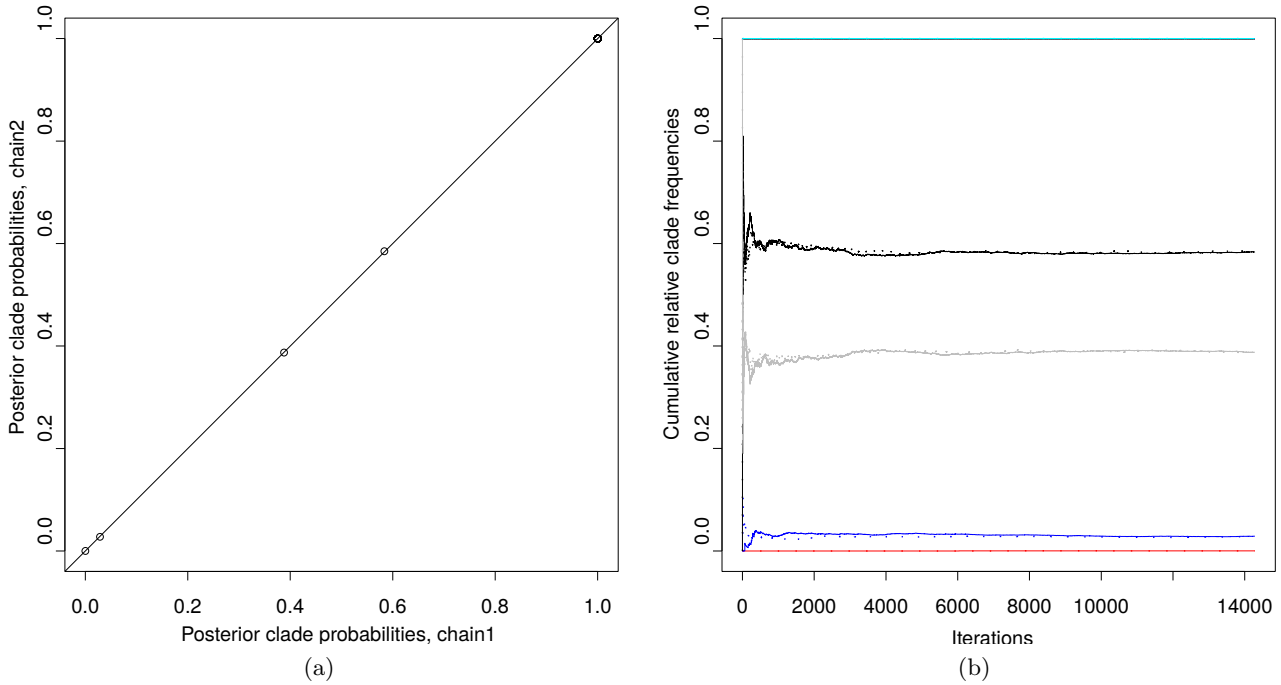
$$\pi(\tau, \boldsymbol{\theta}, \boldsymbol{\ell}|D) = \frac{\pi(\tau|\boldsymbol{\ell}, \boldsymbol{\theta}) \times \pi(\boldsymbol{\ell}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times f(D|\tau, \boldsymbol{\theta}, \boldsymbol{\ell})}{\sum_{i=1}^{T} \int_{\boldsymbol{\ell}} \int_{\boldsymbol{\theta}} \pi(\tau_i|\boldsymbol{\ell}, \boldsymbol{\theta}) \times \pi(\boldsymbol{\ell}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times f(D|\tau_i, \boldsymbol{\theta}, \boldsymbol{\ell}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\ell}},$$

where τ is the tree topology, $\boldsymbol{\ell}$ are the branch lengths, $\boldsymbol{\theta}$ includes all the parameters of the substitution model, and D is the sequence data. Often the parameter of interest is only the tree τ. In this case the parameters of the substitution model and the branch lengths are treated as nuisance parameters by integrating out. Thus, the probability of the tree

given the sequence data is

$$\pi(\tau|D) = \frac{\int_{\boldsymbol{\ell}} \int_{\boldsymbol{\theta}} \pi(\tau|\boldsymbol{\ell},\boldsymbol{\theta}) \times \pi(\boldsymbol{\ell}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times f(D|\tau,\boldsymbol{\theta},\boldsymbol{\ell}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\ell}}{\sum_{i=1}^{T} \int_{\boldsymbol{\ell}} \int_{\boldsymbol{\theta}} \pi(\tau_i|\boldsymbol{\ell},\boldsymbol{\theta}) \times \pi(\boldsymbol{\ell}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times f(D|\tau_i,\boldsymbol{\theta},\boldsymbol{\ell}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\ell}}.$$

The denominator is analytically intractable, since it involves an integral over all branch lengths $\boldsymbol{\ell}$ and all the substitution model parameters $\boldsymbol{\theta}$ for every topology τ. The MCMC technique is used in order to sample from the posterior distribution of the trees. Convergence in the space of phylogenetic trees is usually assessed by comparing the split (branch point where a single lineage evolved into a distinct new one) frequencies between chains initialised at different starting points. An additional diagnostic is analysing the plots of the posterior probabilities of the clades and the cumulative relative frequencies (Heaps *et al.*, 2014). A straight line pattern on a scatter plot of the posterior probabilities of the clades from the two chains suggests convergence has been achieved (Figure 2.12a). Likewise, if the plots of cumulative clade frequencies from both chains are approaching the same fixed values this further indicates convergence (Figure 2.12b).



(a) (b)

Figure 2.12: Graphical convergence diagnostic of the two chains. The plots indicate that the chains have reached convergence. (a) Scatter plot of the posterior probabilities of the clades from the two chains. (b) Plot of cumulative relative clade frequencies. Solid and dotted lines of each colour represent the frequency of the same clade for two different chains.

Chapter 3

Non-reversible substitution models

This chapter focuses on two non-reversible substitution models. Both models incorporate hierarchical priors which are centred on a standard reversible rate matrix but allow non-reversible perturbations of the individual elements. The two models differ in the structure of the perturbation.

A non-reversible model which is centered on a standard reversible rate matrix has been explored previously (Huelsenbeck *et al.*, 2002). However, that study utilised independent uniform distribution for each of the off-diagonal elements of the rate matrix Q, that is $q_{ij} \sim \mathrm{U}(0.001, 100)$, $i \neq j$. Here, we incorporate prior structure for the instantaneous rate matrix, thus adding a biological interpretation to substitution rates. Additionally, in Huelsenbeck *et al.* (2002) the unrooted topology was fixed and the numbers of taxa was small (eight taxa in the simulation study and five taxa in the analysis of the real data). Here, we do not fix the unrooted topology and perform our analysis on larger numbers of taxa (30 taxa in the simulation study and up to 36 taxa in the analysis of the real data). Some work for this chapter appears in Williams *et al.* (2015).

3.1 One component model

3.1.1 Model description

This model, henceforth called the NR model, is based on a log-normal perturbation of the off-diagonal elements of the rate matrix of the HKY85 model. Let $Q^H = (q_{ij}^H)$ be the rate matrix of the HKY85 model, which is characterised by a composition vector

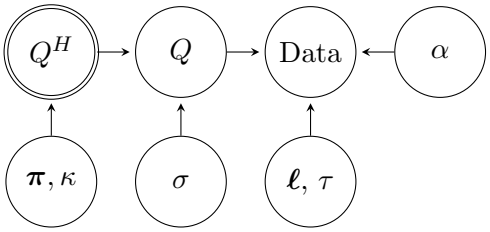
$\boldsymbol{\pi} = (\pi_A,\ \pi_G,\ \pi_C,\ \pi_T)$ and transition-transversion rate ratio κ as follows:

$$Q^H = \begin{pmatrix} \star & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & \star & \pi_C & \pi_T \\ \pi_A & \pi_G & \star & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & \star \end{pmatrix}.$$

Here the symbol $\star$ is used to indicate that the diagonal elements are specified such that every row sums to zero. Let $Q = (q_{ij})$ denote the rate matrix of the NR model. Working element-wise on a log-scale, the off-diagonal elements of the rate matrix of the NR model can be expressed as, for $i \neq j$

$$\log q_{ij} = \log q_{ij}^H + \epsilon_{ij},$$

where the ϵ_{ij} are independent $N(0, \sigma^2)$ quantities. Here the perturbation standard deviation σ represents the extent to which Q departs from a HKY85-structure: the larger its value, the greater the degree of departure. The parameter σ is treated as an unknown quantity whose value we learn about during the analysis. Let us denote by $\boldsymbol{\pi}_Q = (\pi_{Q,A},\ \pi_{Q,G},\ \pi_{Q,C},\ \pi_{Q,T})$ the theoretical stationary distribution which can be obtained from the rate matrix Q (i.e. $\boldsymbol{\pi}_Q Q = \mathbf{0}$). We note that $\boldsymbol{\pi}_Q$ is not the same as $\boldsymbol{\pi}$, the stationary distribution of the underlying HKY85 model. The structure of the hierarchical model for sequence data can therefore be represented through the following directed acyclic graph (DAG):



Here, the data depend on the non-reversible rate matrix Q, the branch lengths $\boldsymbol{\ell}$, the tree topology τ and the across site heterogeneity parameter α. They are conditionally independent of $\boldsymbol{\pi}$, κ and σ given Q.

3.1.2 Likelihood

As shown in Section 2.1.5, under reversible models the likelihood of the data does not depend on the root position. Since the NR model relaxes the reversibility condition, it gives rise to a likelihood function which depends on the position of the root. The NR model is across-branch homogeneous (the same rate matrix is applied to every edge). Processes at different sites are assumed to be independent, but each site i has its own rate of evolution r_i which is modelled by a gamma distribution with mean equal to 1 (Yang,

1993). For computational convenience we approximate the continuous gamma $\mathrm{Ga}(\alpha, \alpha)$ distribution with a discrete $\mathrm{Ga}(\alpha, \alpha)$ distribution with four categories (Yang, 1994), as described in Section 2.5.

In order to calculate the likelihood we first need to calculate transition probability matrices for each branch of the tree. Recall that the transition probability matrix is calculated by using the diagonal form of Q: $P(\ell) = \exp(Q\ell) = \exp(U \times D\ell \times U^{-1}) = U \times \exp(D\ell) \times U^{-1}$, where D is a diagonal matrix of the eigenvalues of Q. In the reversible case, all eigenvalues are real and the calculation is straightforward, i.e. $\exp(D\ell) = \mathrm{diag}(\mathrm{e}^{\lambda_1 \ell}, \ldots, \mathrm{e}^{\lambda_n \ell})$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of Q. However, relaxing the reversibility condition gives rise to rate matrices whose eigenvalues might be complex (Sinclair & Jerrum, 1989). Complex eigenvalues appear in a conjugate pair $\lambda \pm \mathrm{i}\mu$, where λ and μ are real numbers, and i is the imaginary unit which satisfies the equation $\mathrm{i}^2 = -1$. Suppose a pair of complex eigenvalues is present. We will show that out of four eigenvalues of Q the other two eigenvalues are real. Recall that the row sum of the rate matrix Q is zero, i.e. $Q\mathbf{1} = \mathbf{0}$. This implies $Q\mathbf{1} = 0 \times \mathbf{1}$, where $\mathbf{1}$ is the eigenvector and 0 is the eigenvalue. Thus by definition one of the eigenvalues of Q is 0, so the remaining eigenvalue has to be real. Therefore it is possible to have at most one pair of complex eigenvalues out of four eigenvalues of Q.

In the programming language Java, which we use to implement our numerical inference algorithm, the *DenseDoubleEigenvalueDecomposition* class for computing eigenvalues and eigenvectors of a real matrix Q returns a matrix of eigenvectors U and a matrix containing eigenvalues D such that $Q = UDU^{-1}$. If all the eigenvalues are real, the matrix D is diagonal and the calculation of $\exp(D\ell)$ is straightforward. However, if a conjugate pair of eigenvalues is present, the matrix D is block-diagonal, where

$$D = \mathrm{diag}(\lambda_1,\ \Lambda_2,\ \lambda_3)$$

and

$$\Lambda_2 = \begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}.$$

In this case $\exp(D\ell) = \mathrm{diag}(\mathrm{e}^{\lambda_1 \ell}, \mathrm{e}^{\Lambda_2 \ell}, \mathrm{e}^{\lambda_3 \ell})$. Therefore, in order to compute $\exp(D\ell)$, we need to compute e^{Λ_2}. For this we will use the diagonal form of Λ_2, i.e. $\Lambda_2 = XVX^{-1}$ where X is a matrix containing the eigenvectors of Λ_2, and V is a diagonal matrix containing

the eigenvalues of Λ_2. First we will find the eigenvalues and the eigenvectors of Λ_2:

$$\begin{aligned}
&\begin{pmatrix} \lambda - x & \mu \\ -\mu & \lambda - x \end{pmatrix} = 0 \\
&(\lambda - x)^2 + \mu^2 = 0 \\
&(\lambda - x)^2 = -\mu^2 \\
&\lambda - x = \pm \mathrm{i}\mu,
\end{aligned}$$

so the eigenvalues of Λ_2 are $x_i = \lambda \pm \mathrm{i}\mu,\ i = 1, 2$. For eigenvalues $x_i,\ i = 1, 2$ the corresponding eigenvectors are $\mathbf{v_i} = (v_{i1},\ v_{i2})^T$, where

$$\begin{aligned}
&\Lambda_2 \mathbf{v_i} = x_i \mathbf{v_i} \\
&(\Lambda_2 - x_i I_2)\mathbf{v_i} = 0
\end{aligned}$$

for $i = 1, 2$. Therefore we need to solve

$$\begin{pmatrix} \lambda - (\lambda + \mathrm{i}\mu) & \mu \\ -\mu & \lambda - (\lambda + \mathrm{i}\mu) \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{3.1}$$

for $\mathbf{v_1} = (v_{11},\ v_{12})^T$ and

$$\begin{pmatrix} \lambda - (\lambda - \mathrm{i}\mu) & \mu \\ -\mu & \lambda - (\lambda - \mathrm{i}\mu) \end{pmatrix} \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{3.2}$$

for $\mathbf{v_2} = (v_{21},\ v_{22})^T$. Solving (3.1) and (3.2) we get

$$\mathbf{v_1} = (1,\ \mathrm{i})^T$$

and

$$\mathbf{v_2} = (1,\ -\mathrm{i})^T.$$

Hence, the diagonal form of Λ_2 can be written as

$$\Lambda_2 = \begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix} \begin{pmatrix} \lambda + \mathrm{i}\mu & 0 \\ 0 & \lambda - \mathrm{i}\mu \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}^{-1}.$$

Since

$$\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}^{-1} = \frac{1}{-\mathrm{i} - \mathrm{i}} \begin{pmatrix} -\mathrm{i} & -1 \\ -\mathrm{i} & 1 \end{pmatrix} = \frac{\mathrm{i}}{2} \begin{pmatrix} -\mathrm{i} & -1 \\ -\mathrm{i} & 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -\mathrm{i} \\ 1 & \mathrm{i} \end{pmatrix},$$

the diagonal form of Λ_2 can be re-written as

$$\Lambda_2 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}\begin{pmatrix} \lambda+\mathrm{i}\mu & 0 \\ 0 & \lambda-\mathrm{i}\mu \end{pmatrix}\begin{pmatrix} 1 & -\mathrm{i} \\ 1 & \mathrm{i} \end{pmatrix}.$$

Now we can take an exponential of Λ_2 in its diagonal form:

$$\begin{aligned}
\exp\left\{\begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}\right\} &= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}\begin{pmatrix} \exp(\lambda+\mathrm{i}\mu) & 0 \\ 0 & \exp(\lambda-\mathrm{i}\mu) \end{pmatrix}\begin{pmatrix} 1 & -\mathrm{i} \\ 1 & \mathrm{i} \end{pmatrix} \\
&= \frac{\exp(\lambda)}{2}\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}\begin{pmatrix} \cos\mu+\mathrm{i}\sin\mu & 0 \\ 0 & \cos\mu-\mathrm{i}\sin\mu \end{pmatrix}\begin{pmatrix} 1 & -\mathrm{i} \\ 1 & \mathrm{i} \end{pmatrix} \\
&= \frac{\exp(\lambda)}{2}\begin{pmatrix} \cos\mu+\mathrm{i}\sin\mu & \cos\mu-\mathrm{i}\sin\mu \\ -\sin\mu+\mathrm{i}\cos\mu & -\sin\mu-\mathrm{i}\cos\mu \end{pmatrix}\begin{pmatrix} 1 & -\mathrm{i} \\ 1 & \mathrm{i} \end{pmatrix} \\
&= \frac{\exp(\lambda)}{2}\begin{pmatrix} 2\cos\mu & 2\sin\mu \\ -2\sin\mu & 2\cos\mu \end{pmatrix} \\
&= \exp(\lambda)\begin{pmatrix} \cos\mu & \sin\mu \\ -\sin\mu & \cos\mu \end{pmatrix}.
\end{aligned}$$

So, if a matrix Q has a pair of complex eigenvalues, its diagonal form is represented in Java as

$$Q = U \times \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & \mu & 0 \\ 0 & -\mu & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_3 \end{pmatrix} \times U^{-1},$$

where $\lambda_2 \pm \mathrm{i}\mu$ is the conjugate pair of complex eigenvalues, and λ_1 and λ_3 are real eigenvalues. The transition probability matrix is thus

$$P(\ell) = \exp(Q\ell) = U \times \exp\begin{pmatrix} \lambda_1\ell & 0 & 0 & 0 \\ 0 & \lambda_2\ell & \mu\ell & 0 \\ 0 & -\mu\ell & \lambda_2\ell & 0 \\ 0 & 0 & 0 & \lambda_3\ell \end{pmatrix} \times U^{-1}$$

$$= U \times \begin{pmatrix} \mathrm{e}^{\lambda_1\ell} & 0 & 0 & 0 \\ 0 & \mathrm{e}^{\lambda_2\ell}\cos\mu\ell & \mathrm{e}^{\lambda_2\ell}\sin\mu\ell & 0 \\ 0 & -\mathrm{e}^{\lambda_2\ell}\sin\mu\ell & \mathrm{e}^{\lambda_2\ell}\cos\mu\ell & 0 \\ 0 & 0 & 0 & \mathrm{e}^{\lambda_3\ell} \end{pmatrix} \times U^{-1}.$$

Note that the block-diagonal matrix D can alternatively be of the form

$$D = \text{diag}(\Lambda_1,\ \lambda_2,\ \lambda_3)$$

or

$$D = \text{diag}(\lambda_1,\ \lambda_2,\ \Lambda_3)$$

in which case the calculation is performed in an analogous way.

3.1.3 Prior

The aim of the analysis is to infer the parameters of the model: the composition vector $\boldsymbol{\pi}$, the transition-transversion rate ratio κ, the perturbation standard deviation σ, the off-diagonal elements of the rate matrix Q, the shape parameter of the gamma distribution for the across site variation α, the branch lengths $\boldsymbol{\ell}$ and the rooted topology τ. We express our uncertainty about these unknown parameters though a prior distribution which is given by

$$\pi(\boldsymbol{\pi}, \kappa, \sigma, Q, \alpha, \boldsymbol{\ell}, \tau) = \pi(\boldsymbol{\pi})\pi(\kappa)\pi(\sigma)\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)\pi(\alpha)\pi(\boldsymbol{\ell})\pi(\tau).$$

Priors for numerical parameters

The composition vector $\boldsymbol{\pi}$ is defined on the four-dimensional simplex, that is, it has four positive elements, constrained to sum to one. We choose to assign it a Dirichlet prior, $\boldsymbol{\pi} \sim \mathcal{D}(\alpha_\pi \boldsymbol{\pi}_0)$, where $\boldsymbol{\pi}_0 = (0.25,\ 0.25,\ 0.25,\ 0.25)$ is the mean and α_π is a concentration parameter (we take $\alpha_\pi = 4$). This prior is exchangeable with respect to the nucleotide labels. We adopt a log-normal prior for the transition-transversion rate ratio $\kappa \sim \text{LN}(\log \kappa_0, \xi^2)$, where $\kappa_0 = 1$ and $\xi = 0.8$. The parameters of the prior for κ represent our belief that the probability of κ exceeding 2 is 0.2, i.e. $\Pr(\kappa < 2) = 0.8$. Our choice of the priors for $\boldsymbol{\pi}$ and κ is governed by the biological opinion about these parameters.

The perturbation parameter σ is assigned an exponential prior $\sigma \sim \text{Exp}(\gamma)$, where the rate $\gamma = 2.3$ reflects our prior belief that the probability of σ exceeding 1 is 0.1, i.e. $\Pr(\sigma < 1) = 0.9$. This choice discourages a stationary distribution $\boldsymbol{\pi}_Q$ in which some characters are heavily favoured over the others. Figure 3.1 shows a boxplot of 1000 samples from the prior for the first element of the stationary distribution for different values of σ. As σ increases, significant support is given to highly biased compositions, and for $\sigma > 1.0$ these are biologically unrealistic.

The branch lengths are assigned independent exponential priors $\ell_i \sim \text{Exp}(\mu)$, where $i = 1, \ldots, k$ and k is the number of edges. The rate μ equals 10, so that $\text{E}(\ell_i) = 0.1$ in keeping with biologists' beliefs about the number of substitutions per site. The shape parameter α is assigned a gamma prior, $\alpha \sim \text{Ga}(10, 10)$, which ensures the expected

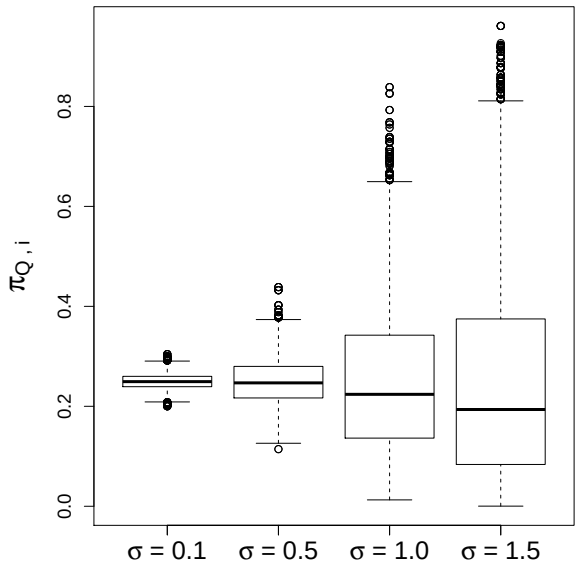


Figure 3.1: Boxplot of the prior for the first element of the stationary distribution for different values of the perturbation standard deviation σ conditional on the rate matrix Q^H (the priors for the rest of the elements of the stationary distribution are the same due to symmetry). Increasing the value of σ clearly increases the spread in the prior for the stationary probabilities.

substitution rate in the $\text{Ga}(\alpha, \alpha)$ model for site-specific substitution rates is modestly concentrated around 1.

Priors for topology

We define a *root type* as the number of species on each side of the root. For example, the root type $1 : (n-1)$ represents a root split on a pendant edge, $2 : (n-2)$ represents a root split between two taxa and all others, etc. A uniform prior over rooted topologies assigns a prior probability of more than 0.5 to root splits of the type $1 : (n-1)$, in other words, to roots on pendant edges (an unrooted tree of n taxa has $2n-3$ branches, so the probability of the root split on pendant edges is $n/(2n-3) > 0.5$). We felt that deeper roots are generally more biologically plausible and should be assigned higher prior mass, whilst still retaining a diffuse initial distribution. We therefore chose to assign the rooted tree topology a prior according to the Yule model (birth process), which assumes that at any given time each of the species is equally likely to undergo a speciation event. This generates a biologically defensible prior in which all root types receive the same prior probability if n is odd, and a near uniform distribution if n is even, but with $n/2 : n/2$ root types receiving half the prior probability of the other root types. This property follows from the fact that if we randomly select one of the two sub-trees incident with the root

of the tree generated under the Yule model, then the number of leaves in this sub-tree is uniformly distributed between 1 and $n-1$, that is $f(i) = 1/(n-1)$, where $f(i)$ is the probability that the number of leaves in the sub-tree is i ($i = 1, \ldots, n-1$).

A *labelled history* is a rooted tree with the internal vertices rank-ordered according to their ages (Yang, 2014). The probability of generating a n-species tree T under the Yule distribution is calculated by dividing the number of labelled histories for the tree T by the total number of all possible labelled histories on n species:

$$\frac{2^{n-1}}{n!}\left(\prod_{v\in T_0}\lambda_v\right)^{-1},$$

where T_0 is the set of interior vertices of the rooted tree T, and λ_v is the number of internal vertices that are descendants of v, or one less than the number of leaves below v (Steel & McKenzie, 2001). This probability depends on the complete rooted topology. As discussed in Section 2.6 we fit our model to data using Bayesian inference via a Metropolis-within-Gibbs sampler. An important step is computation of the prior ratio. Therefore if we use a Yule prior, we have to re-calculate the prior probability of the tree at every MCMC iteration.

To save computational time, we therefore additionally introduce an approximation to the Yule prior, which we term the *structured uniform prior*, which assigns equal prior probability to all root-types, and also equal prior probability to all rooted trees within each root-type. If we denote by k the number of taxa on one side of the root, then $n-k$ is the number of taxa on the other side of the root. The probability of generating a n-species tree T under the structured uniform prior is

$$\frac{1}{t\binom{n}{k}N_kN_{n-k}},$$

where

$$\frac{1}{t} = \begin{cases} n/2, & \text{if } n \text{ is even} \\ (n-1)/2, & \text{if } n \text{ is odd} \end{cases}$$

is the probability of choosing a root-type $k:(n-k)$,

$$\frac{1}{\binom{n}{k}N_kN_{n-k}}$$

is the probability of choosing a rooted tree given a root-type $k:(n-k)$, and

$$N_i = \frac{(2i-2)!}{2^{i-1}(i-1)!}$$

is the number of rooted trees with i taxa. Computationally, this prior is more convenient than the Yule prior because its mass function is independent of the particular rooted topology given its root type and so only considers the root type. It also has the advantage of being uniform on root types for all n.

3.1.4 Posterior inference via MCMC

According to Bayes' theorem, the posterior distribution is proportional to the prior times the likelihood and is given by

$$\pi(\boldsymbol{\pi}, \kappa, \sigma, Q, \alpha, \boldsymbol{\ell}, \tau | D) \propto \pi(Q|\boldsymbol{\pi}, \kappa, \sigma) \times \pi(\boldsymbol{\pi}, \kappa, \sigma, \alpha, \boldsymbol{\ell}, \tau) \times \pi(D|Q, \alpha, \boldsymbol{\ell}, \tau).$$

This distribution is analytically intractable, therefore we utilise Markov chain Monte Carlo (MCMC) methods, specifically a Metropolis-within-Gibbs sampling scheme to generate dependent samples from the posterior. At each iteration of the MCMC algorithm the following steps are performed:
(a) update the parameters of the substitution model, i.e. $\boldsymbol{\pi}, \kappa, \sigma, Q, \alpha$;
(b) update the branch lengths $\boldsymbol{\ell}$ and the topology τ.

In step (a) we update the parameters one at a time, sweeping through a Dirichlet random walk proposal for $\boldsymbol{\pi}$ and log-normal random walk proposals for the rest of the parameters as follows:

(i) Metropolis-Hastings step for the composition vector $\boldsymbol{\pi}$:

Prior: $\boldsymbol{\pi} \sim \mathcal{D}(\alpha_\pi \boldsymbol{\pi}_0)$, $\alpha_\pi = 4,\ \boldsymbol{\pi}_0 = (0.25,\ 0.25,\ 0.25,\ 0.25)$.
Proposal: $\boldsymbol{\pi}' \sim \mathcal{D}(a_\pi \boldsymbol{\pi})$, where $\boldsymbol{\pi}$ is the current value, a_π is a tuning parameter.
Acceptance probability:

$$\min\left\{1, \frac{\pi(\boldsymbol{\pi}')}{\pi(\boldsymbol{\pi})} \times \frac{q(\boldsymbol{\pi}|\boldsymbol{\pi}')}{q(\boldsymbol{\pi}'|\boldsymbol{\pi})} \times \frac{\pi(Q|\boldsymbol{\pi}', \kappa, \sigma)}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \times \frac{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)}\right\}.$$

After cancelling the term $\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)$ in the numerator and denominator, the acceptance probability takes the form $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\pi}')}{\pi(\boldsymbol{\pi})} \times \frac{q(\boldsymbol{\pi}|\boldsymbol{\pi}')}{q(\boldsymbol{\pi}'|\boldsymbol{\pi})} \times \frac{\pi(Q|\boldsymbol{\pi}', \kappa, \sigma)}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \\ &= \prod_{i=1}^{4} \frac{\Gamma(a_\pi \pi_i)}{\Gamma(a_\pi \pi_i')} \pi_i^{(a_\pi \pi_i' - \alpha_\pi \pi_{0i})} \pi_i'^{(\alpha_\pi \pi_{0i} - a_\pi \pi_i)} \\ &\quad \times \exp\left[\frac{1}{2\sigma^2} \sum_{i \neq j} \left\{(\log q_{ij}^H)^2 - (\log q_{ij}^{H'})^2 + 2 \log q_{ij} (\log q_{ij}^{H'} - \log q_{ij}^H)\right\}\right], \end{aligned}$$

and the $q_{ij}^{H'}$ are the off-diagonal elements of the HKY85 rate matrix computed with $\boldsymbol{\pi}'$.

In practice using a Dirichlet random walk proposal can cause computational problems for low values of the concentration parameter, because the mean of the proposal distribution can be close to zero. In this case the variance of the proposal distribution is also close to zero, thus resulting in the sampler getting stuck at values close to zero. In order to avoid this problem we use a Dirichlet random walk proposal with a nudge to keep the mean away from zero (Germain, 2010; Loza-Reyes *et al.*, 2014). We choose a value of $\delta = 0.005$ for the nudge, as suggested in Germain (2010). The proposal then has mean with elements, for $i = 1, \ldots, 4$

$$\pi_{\delta,i} = \frac{a_\pi \pi_i + \delta}{a_\pi + 4\delta}.$$

(ii) Metropolis-Hastings step for the transition-transversion rate ratio κ:

Prior: $\kappa \sim \text{LN}(\log \kappa_0, \xi^2)$, $\kappa_0 = 1,\ \xi = 0.8$.
Proposal: $\kappa' \sim \text{LN}(\log \kappa, a_\kappa^2)$, where κ is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\kappa')}{\pi(\kappa)} \times \frac{q(\kappa|\kappa')}{q(\kappa'|\kappa)} \times \frac{\pi(Q|\boldsymbol{\pi}, \kappa', \sigma)}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \times \frac{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \frac{\pi(\kappa')}{\pi(\kappa)} \times \frac{q(\kappa|\kappa')}{q(\kappa'|\kappa)} \times \frac{\pi(Q|\boldsymbol{\pi}, \kappa', \sigma)}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \\ &= \exp\left[\frac{1}{2\xi^2}\left\{(\log\kappa)^2 - (\log\kappa')^2 + 2\log\kappa_0(\log\kappa' - \log\kappa)\right\}\right] \\ &\qquad \times \exp\left[\frac{1}{2\sigma^2}\sum_{i\neq j}\left\{(\log q_{ij}^H)^2 - (\log q_{ij}^{H'})^2 + 2\log q_{ij}(\log q_{ij}^{H'} - \log q_{ij}^H)\right\}\right], \end{aligned}$$

and $q_{ij}^{H'}$ are the off-diagonal elements of the HKY85 rate matrix computed with κ'.

(iii) Metropolis-Hastings step for the perturbation standard deviation σ:

Prior: $\sigma \sim \text{Exp}(\gamma)$, $\gamma = 2.3$.
Proposal: a mixture of
(1) random walk proposal $\sigma' \sim \text{LN}(\log\sigma, a_\sigma^2)$, where σ is the current value;
(2) independence sampler proposal $\sigma' \sim \text{Exp}(\gamma)$.
At every iteration of the MCMC algorithm a proposal is chosen uniformly from the two choices above.

Acceptance probability for proposal (1): $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\sigma')}{\pi(\sigma)} \times \frac{q(\sigma|\sigma')}{q(\sigma'|\sigma)} \times \frac{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma')}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \times \frac{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \frac{\pi(\sigma')}{\pi(\sigma)} \times \frac{q(\sigma|\sigma')}{q(\sigma'|\sigma)} \times \frac{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma')}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \\ &= \left(\frac{\sigma}{\sigma'}\right)^{K-1} \times \exp\left[\gamma(\sigma - \sigma') + \frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{1}{\sigma'^2}\right) \sum_{i \neq j} \left\{(\log q_{ij} - \log q_{ij}^H)^2\right\}\right], \end{aligned}$$

and K is the number of the off-diagonal elements of the rate matrix Q ($K = 12$ for DNA data).

Acceptance probability for proposal (2): $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\sigma')}{\pi(\sigma)} \times \frac{q(\sigma)}{q(\sigma')} \times \frac{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma')}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \times \frac{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \frac{\pi(\sigma')}{\pi(\sigma)} \times \frac{q(\sigma)}{q(\sigma')} \times \frac{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma')}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \\ &= \left(\frac{\sigma}{\sigma'}\right)^{K} \times \exp\left[\frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{1}{\sigma'^2}\right) \sum_{i \neq j} \left\{(\log q_{ij} - \log q_{ij}^H)^2\right\}\right], \end{aligned}$$

and K is the number of the off-diagonal elements of the rate matrix Q.

The reason we use a mixture of proposals rather than a single proposal is that in practice using a log-normal random walk proposal distribution for σ often causes the MCMC sampler getting stuck in local maxima, i.e. regions where there exists a value σ for which the likelihood of the data is higher than for the values within the close neighbours of σ, but lower than for the neighbours of σ which are further away. In order to avoid this problem, we employ a mixture of proposals with two components: a random walk proposal and an independence sampler proposal (Tierney, 1994). The independence sampler proposes a new value of the parameter that is independent of the current value. We use the exponential prior distribution for σ as a kernel for the independence sampler, which has bigger variance than the log-normal random walk proposal. Hence, the sampler has a better potential of leaving local maxima, and allows the MCMC to better explore the space of σ. Using a mixture distribution where components are a log-normal random walk and an independence sampler with an exponential kernel is equivalent to alternating between small and big moves in order to improve the mixing of the chain.

(iv) Metropolis-Hastings steps for the off-diagonal elements of the rate matrix Q consist of a sweep through all of the off-diagonal elements q_{ij}:

Prior for each element: $q_{ij} \sim \text{LN}(\log q_{ij}^H, \sigma^2)$.
Proposal for each element: $q'_{ij} \sim \text{LN}(\log q_{ij}, a_q^2)$, where q_{ij} is the current value.
Acceptance probability for each element: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(Q'|\boldsymbol{\pi}, \kappa, \sigma)}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \times \frac{q(q_{ij}|q'_{ij})}{q(q'_{ij}|q_{ij})} \times \frac{\pi(D|Q', \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \exp\left[\frac{1}{2\sigma^2}\sum_{i\neq j}\left\{(\log q_{ij})^2 - (\log q'_{ij})^2 + 2\log q_{ij}^H(\log q'_{ij} - \log q_{ij})\right\}\right] \\ &\quad \times \frac{\pi(D|Q', \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

(v) Metropolis-Hastings step for the gamma shape heterogeneity parameter α:

Prior: $\alpha \sim \text{Ga}(s, r)$, $s = 10, r = 10$.
Proposal: $\alpha' \sim \text{LN}(\log \alpha,\, a_\alpha^2)$, where α is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\alpha')}{\pi(\alpha)} \times \frac{q(\alpha|\alpha')}{q(\alpha'|\alpha)} \times \frac{\pi(D|Q, \alpha', \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \left(\frac{\alpha'}{\alpha}\right)^s \exp\left\{r(\alpha - \alpha')\right\} \times \frac{\pi(D|Q, \alpha', \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

Step (b) consists of a series of Metropolis-Hastings steps to update each branch length one at a time and then updating the rooted topology and branch lengths (in a joint move) through three types of proposal: nearest-neighbour interchange (NNI), sub-tree prune and regraft (SPR) (Allen & Steeel, 2001; Yang, 2006), and a proposal which moves the root (Heaps *et al.*, 2014).

(i) Metropolis-Hastings step for the branch lengths $\boldsymbol{\ell}$:

Prior for each branch length: $\ell_i \sim \text{Exp}(\mu)$, $\mu = 10$.
Proposal for each branch length: a mixture of
(1) random walk proposal $\ell'_i \sim \text{LN}(\log \ell_i, a_\ell^2)$, where ℓ_i is the current value;
(2) independence sampler proposal $\ell'_i \sim \text{Exp}(\mu)$.
At every iteration of the MCMC algorithm a proposal is chosen uniformly from the two choices above.

Acceptance probability for proposal (1): $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\ell}')}{\pi(\boldsymbol{\ell})} \times \frac{q(\ell_i|\ell_i')}{q(\ell_i'|\ell_i)} \times \frac{\pi(D|Q,\alpha,\boldsymbol{\ell}',\tau)}{\pi(D|Q,\alpha,\boldsymbol{\ell},\tau)} \\ &= \frac{\ell_i'}{\ell_i} \times \exp(\ell_i - \ell_i') \times \frac{\pi(D|Q,\alpha,\boldsymbol{\ell}',\tau)}{\pi(D|Q,\alpha,\boldsymbol{\ell},\tau)}. \end{aligned}$$

Acceptance probability for proposal (2): $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\ell}')}{\pi(\boldsymbol{\ell})} \times \frac{q(\ell_i)}{q(\ell_i')} \times \frac{\pi(D|Q,\alpha,\boldsymbol{\ell}',\tau)}{\pi(D|Q,\alpha,\boldsymbol{\ell},\tau)} \\ &= \frac{\pi(D|Q,\alpha,\boldsymbol{\ell}',\tau)}{\pi(D|Q,\alpha,\boldsymbol{\ell},\tau)}. \end{aligned}$$

A mixture of two proposals for branch lengths in which the innovation variance of one component depends on the current branch length and the innovation variance of the other does not, has been shown to produce better mixing than using a single proposal. While the proposal in which the innovation variance depends on the current value results in better mixing for short branches, the proposal in which the innovation variance does not depend on the current value produces better mixing for long branches. Thus alternating the two types of proposals results in better mixing of the chain (Loza-Reyes *et al.*, 2014).

(ii) Metropolis-Hastings step for the NNI move:

The nearest-neighbour interchange (NNI) algorithm is a topological rearrangement of a tree which works by swapping two sub-trees on the two sides of a branch. First an internal branch e is selected uniformly at random (excluding the two branches adjacent to the root vertex). Let us denote the vertex on e which is closest to the root by v_0, and the vertex which is closest to the leaves by v. We denote the two sub-trees descended from v by T_1 and T_2, and the sub-tree descended from v_0 by T_0. In the NNI move either sub-tree T_1 or T_2 is replaced with the sub-tree T_0, as illustrated in Figure 3.2 (the probability of choosing either T_1 or T_2 is $1/2$). Thus a single NNI move can result in one of the two possible trees as shown in Figure 3.3. The process creates a new branch e' which replaces the branch e. The length of e' is proposed using a log-normal random walk proposal centred on the length of e. It is worth noting that the root on the new tree remains unchanged since the swapped sub-trees are descended from an edge not adjacent to the root. The acceptance

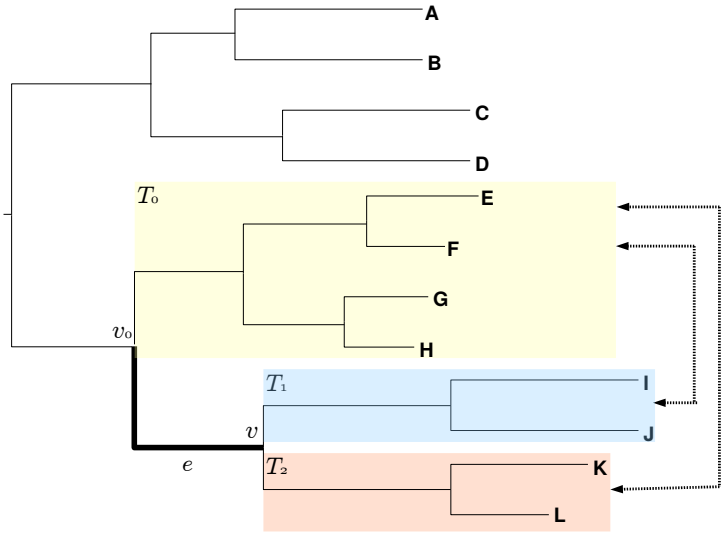


Figure 3.2: An illustration of the NNI move. The internal edge e is chosen uniformly at random from the set of internal edges not adjacent to the root. During the move, either sub-tree T_1 or T_2 descended from the vertex v is interchanged with the sub-tree T_0 descended from the vertex v_0.

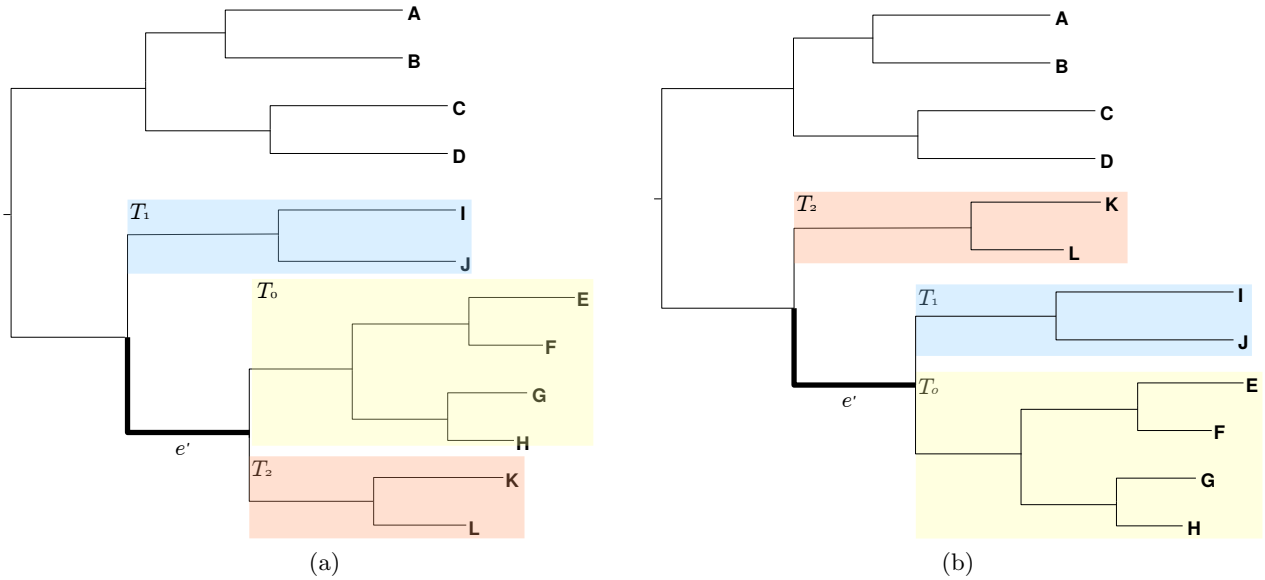


Figure 3.3: Two possible trees resulting from the NNI move illustrated in Figure 3.2. In (a) the sub-tree T_1 is interchanged with the sub-tree T_0. In (b) the sub-tree T_2 is interchanged with the sub-tree T_0. The length of the branch e' is proposed using the log-normal random walk proposal centred on the length of e from the original tree. The root of the new trees remains unchanged.

probability of the NNI move is $\min\{1, A\}$, where

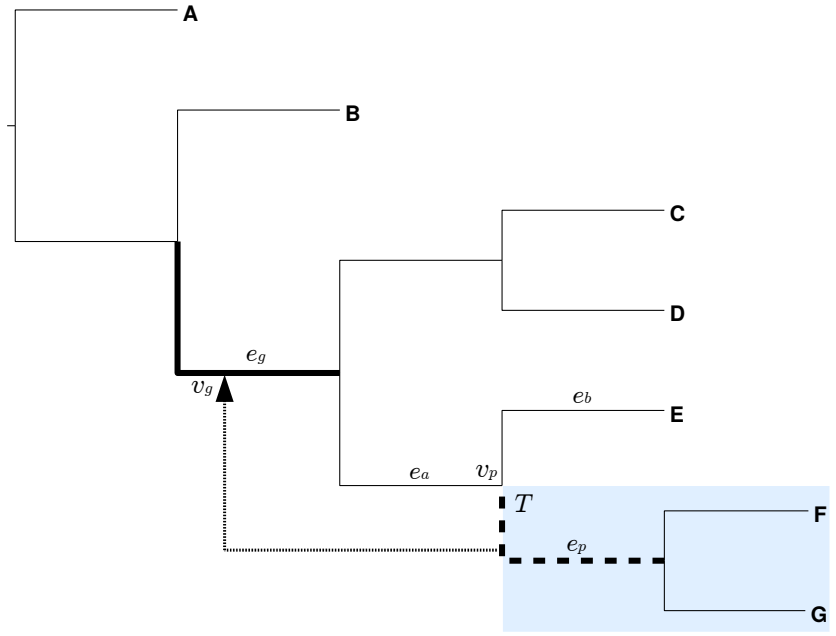
$$\begin{aligned} A &= \frac{\pi(\tau')}{\pi(\tau)} \times \frac{\pi(\boldsymbol{\ell}')}{\pi(\boldsymbol{\ell})} \times \frac{q(\ell_e|\ell_{e'})}{q(\ell_{e'}|\ell_e)} \times \frac{\pi(D|Q,\alpha,\boldsymbol{\ell}',\tau)}{\pi(D|Q,\alpha,\boldsymbol{\ell},\tau)} \\ &= \frac{\pi(\tau')}{\pi(\tau)} \times \exp\{\mu(\ell_e - \ell_{e'})\} \times \frac{\ell_{e'}}{\ell_e} \times \frac{\pi(D|Q,\alpha,\boldsymbol{\ell}',\tau)}{\pi(D|Q,\alpha,\boldsymbol{\ell},\tau)}. \end{aligned}$$

In practice the acceptance rate in this move is very small (around 1%), and the proposal in this move cannot be tuned to achieve the desirable acceptance rate. Thus the posterior samples have high autocorrelation, which causes poor mixing of the chain. It is possible to reduce autocorrelation by performing multiple NNI moves at each MCMC iteration.

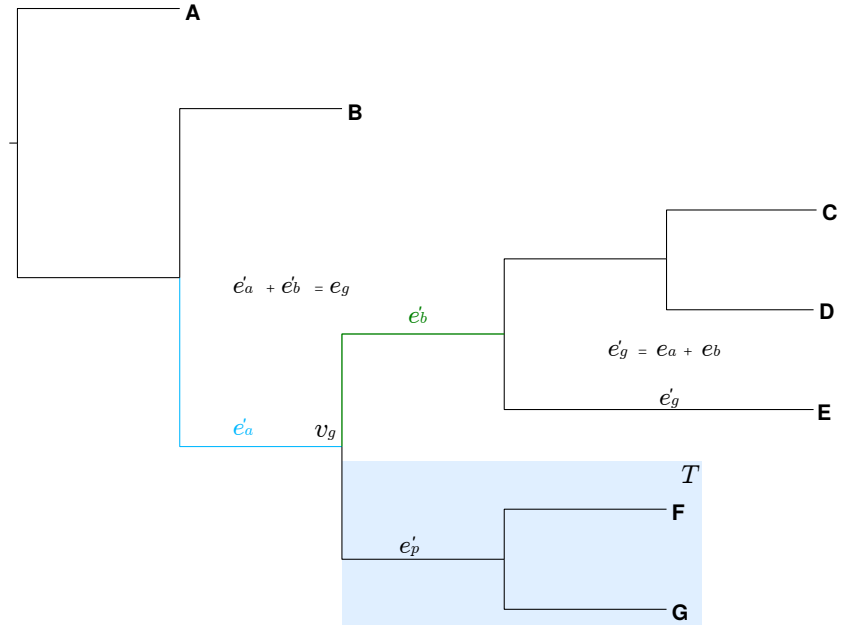
(iii) Metropolis-Hastings step for the SPR move:

Sub-tree pruning and regrafting (SPR) works by pruning a sub-tree and reattaching it to a different branch. The move is illustrated in Figure 3.4. As in the NNI move, we first select uniformly at random an internal edge e_p which is not adjacent to the root. We also select uniformly at random another internal edge e_g which is not adjacent to either e_p or the root. Denote by v_p the vertex closest to the root on the edge e_p. Denote by e_a and e_b the edges containing the vertex v_p. Denote by T the tree evolving from the vertex v_p (including the edge e_p). The SPR move consists of pruning the tree T and reattaching it to a point v_g on the edge e_g (Figure 3.4a). The reattachment divides the edge e_g into two edges: e'_a and e'_b thus introducing a new vertex (v_g) shared by the two newly created edges. The vertex v_p disappears such that the edges e_a and e_b are merged to form a new edge e'_g (Figure 3.4b). The lengths of the edges e'_a and e'_b are proposed as follows: we first sample a random variable $u \sim \text{Beta}(2,2)$, and we set the length of e'_a to be proportional to the value of u, that is $\ell_{e'_a} = u \times \ell_{e_g}$, where ℓ_i is a length of an edge i. The length of the edge e'_b is then set such that the overall branch length is preserved, that is $\ell_{e'_b} = (1-u) \times \ell_{e_g}$. The parameters of the beta distribution of the random variable u are chosen such that $\text{E}(u) = 0.5$, that is the regrafting point is centred on the middle of the edge e_g. The acceptance probability of the SPR move is $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\tau')}{\pi(\tau)} \times \frac{\pi(\boldsymbol{\ell}')}{\pi(\boldsymbol{\ell})} \times \frac{q(w)}{q(u)} \times \left|\frac{\partial(\ell_{e'_a}, \ell_{e'_b}, \ell_{e'_g}, w)}{\partial(\ell_{e_a}, \ell_{e_b}, \ell_{e_g}, u)}\right| \times \frac{\pi(D|Q,\alpha,\boldsymbol{\ell}',\tau')}{\pi(D|Q,\alpha,\boldsymbol{\ell},\tau)} \\ &= \frac{\pi(\tau')}{\pi(\tau)} \times \frac{\pi(\boldsymbol{\ell}')}{\pi(\boldsymbol{\ell})} \times \frac{w(1-w)}{u(1-u)} \times \frac{\ell_{e_g}}{\ell_{e_a} + \ell_{e_b}} \times \frac{\pi(D|Q,\alpha,\boldsymbol{\ell}',\tau')}{\pi(D|Q,\alpha,\boldsymbol{\ell},\tau)}, \end{aligned} \tag{3.3}$$



(a)



(b)

Figure 3.4: An illustration of the SPR move. (a) During the move, the edge e_p (dashed line) and the tree T evolving from it are pruned and reattached to the edge e_g. The attachment point v_g is chosen by dividing the edge e_g using a random variable drawn from Beta$(2, 2)$. (b) As a result of the move, the vertex v_p disappears, such that the edges e_a and e_b are merged to form a new edge e'_g. The grafting edge e_g is split into two new edges e'_a and e'_b by a new vertex v_g which is formed after reattaching the sub-tree T to e_g.

$w = \ell_{e_a}/(\ell_{e_a} + \ell_{e_b})$ is the auxiliary variable for the reverse move, and

$$\left| \frac{\partial(\ell_{e'_a}, \ell_{e'_b}, \ell_{e'_g}, w)}{\partial(\ell_{e_a}, \ell_{e_b}, \ell_{e_g}, u)} \right| = \frac{\ell_{e_g}}{\ell_{e_a} + \ell_{e_b}}$$

is the Jacobian (Blanquart & Lartillot, 2006). As in the case of the NNI move, and for the same reason, the acceptance rate in the SPR move is small (around 1%).

(iv) Metropolis-Hastings step for the root move:

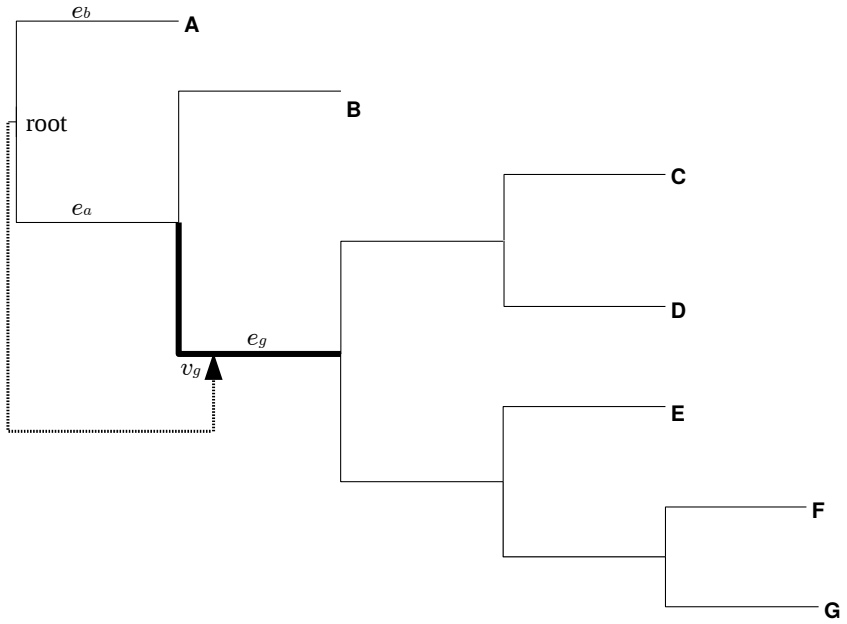
This move is similar to the SPR move. A new root is proposed by selecting a branch e_g uniformly at random from the set of branches not adjacent to the root (Figure 3.5a). The new root is created by inserting a degree two vertex v_g on the branch e_g. Thus the branch e_g is divided into two sub-branches: e'_a and e'_b (Figure 3.5b). The lengths of these sub-branches are proposed using a random variable from the Beta$(2, 2)$ distribution similarly to the SPR move. The existing root vertex disappears such that the two edges e_a and e_b are merged to create a new edge e'_g. Because of the similarity with the SPR move, the acceptance probability of the root move is calculated similarly, according to equation (3.3).

3.1.5 Simulation study

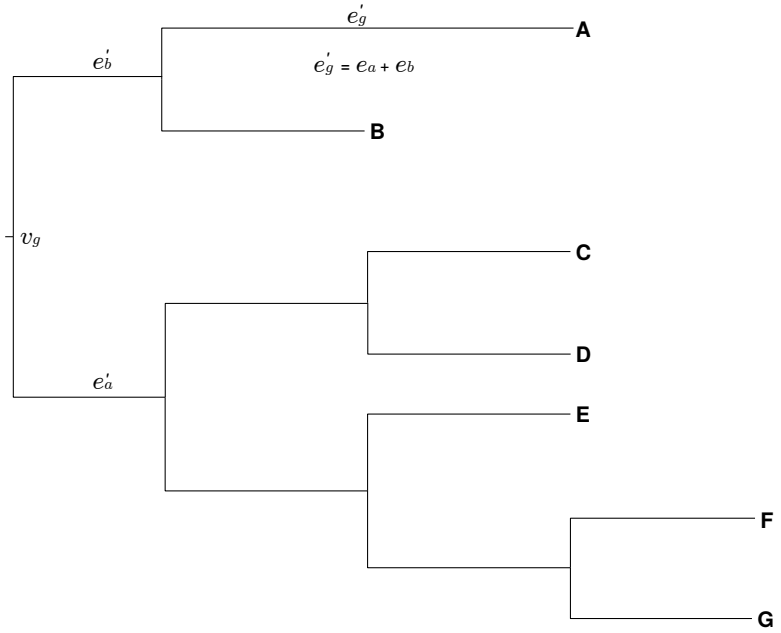
The simulations are divided into two independent blocks. The first block of the simulations aims to investigate root inference for data simulated with different levels of non-reversibility under a random rooted tree. The second block of the simulations explores the effect of different rooted topologies and different branch lengths on the root inference.

Block One

Here we explore the shape of the posterior when the model is fitted to data that contain different levels of non-reversibility. The tree used to simulate the data is a random 30-taxon tree (generated under the Yule birth process), with the branch lengths simulated from Ga$(2, 20)$. The lengths of the branches adjacent to the root are simulated from Ga$(1, 20)$ such that the combined lengths of these two branches will correspond to Ga$(2, 20)$ (Figure 3.6). Since the tree was generated under the Yule birth process we expect both the Yule prior and the structured uniform prior to assign a lot of support to the true root split. Therefore, if we analyse data simulated under this tree, the high prior support for the true root split may be reflected in the posterior, in spite of the information from the data, whose effect we are investigating. In order to perform a more objective experiment, we therefore reroot the tree such that the new root is not favoured by the prior. First, we investigate



(a)



(b)

Figure 3.5: An illustration of the root move. (a) During the move, a new root is created by inserting a degree two vertex v_g on the branch e_g. (b) As a result of the move, the new root v_g divides the branch e_g into two sub-branches: e'_a and e'_b. The existing root vertex disappears such that the two edges e_a and e_b are merged to create a new edge e'_g.

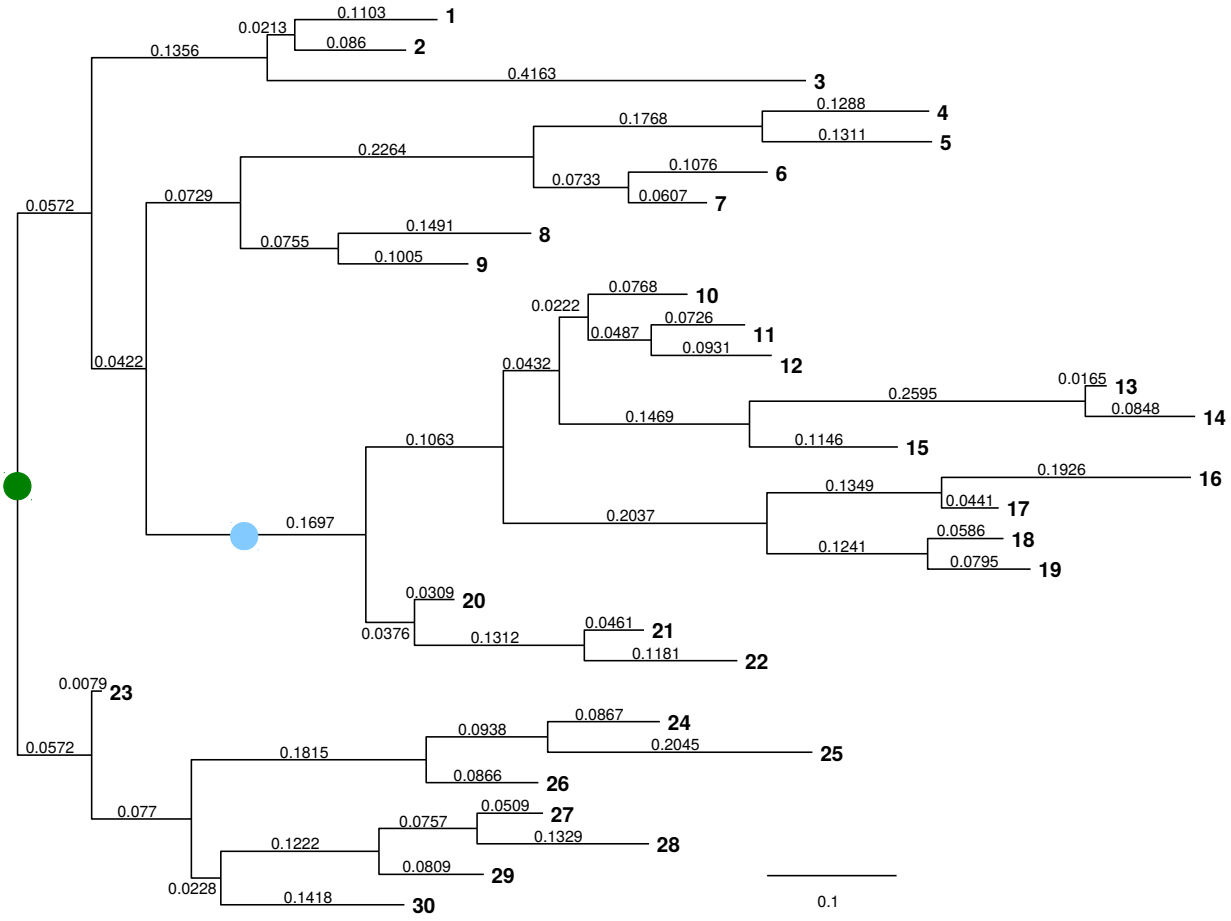
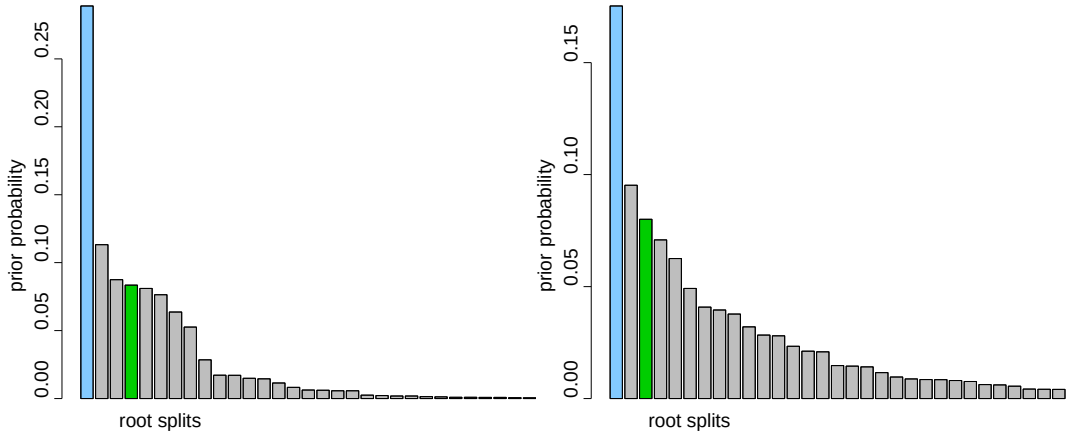


Figure 3.6: Rooted random 30-taxon tree generated under the Yule birth model used to simulate the data in the first block of the simulations. The blue circle maps the branch which is preferred by the topological priors to place the root on. The green circle maps the alternative root split having much lower prior probability. The data were simulated under the tree rooted on the branch mapped with the green circle.

the prior distribution of the root splits conditional on the true unrooted topology and the true values of the branch lengths (Figure 3.7). Indeed, both priors favour the original root split represented by the blue bar in Figure 3.7 (mapped with the blue circle in Figure 3.6). The green bar corresponding to the new root split has much lower prior probability than the original root split. Therefore we simulate the data under the rerooted tree, such that the model will have to extract information from the data rather than rely heavily on the prior information.



(a) Yule distribution on the root splits.

(b) Structured uniform distribution on the root splits.

Figure 3.7: Prior distribution of the root splits conditional on the unrooted topology and branch lengths in Figure 3.6 for (a) the Yule prior; (b) the structured uniform prior. Different bars on the plots represent different root splits on the prior distribution of trees (ordered by prior probabilities). On each plot the blue bar corresponds to the original root split, the green bar correspond to the alternative root split the data were simulated with (both root splits are mapped in Figure 3.6).

In order to simulate the alignments, we first fixed the underlying reversible HKY85 rate matrix (Q^H matrix) using the values $\boldsymbol{\pi} = (0.25,\ 0.25,\ 0.25,\ 0.25)$ and $\kappa = 2$. Then we applied a log-normal perturbation to the Q^H matrix to obtain the non-reversible Q matrix to simulate the data from. To investigate the effect of different levels of non-reversibility, five different values of the perturbation standard deviation were used to simulate the data: $\sigma = 0,\ 0.05,\ 0.1,\ 0.2,\ 0.3$. For each value of the perturbation standard deviation nine different data sets of length 2000 sites were simulated, the first five having different rate matrices (data sets 1 - 5), and the last five having the same rate matrix (data sets 5 - 9). Thus the former five data sets have different stationary distributions, while the latter five data sets have the same stationary distribution. This type of alignment simulation allows us to investigate different sources of variability in the data. All the alignments were simulated using a gamma shape heterogeneity parameter simulated from $\text{Ga}(10, 10)$. Note that the case of $\sigma = 0$ corresponds to the reversible HKY85 model. The other values of σ were chosen so that the prior for the stationary distribution induced by the log-normal

perturbation would be in the range of values estimated for real data; as σ increases, significant support is given to highly biased compositions, and for $\sigma > 0.3$ these are biologically unrealistic (Figure 3.8). The simulation results are based on (almost) un-autocorrelated

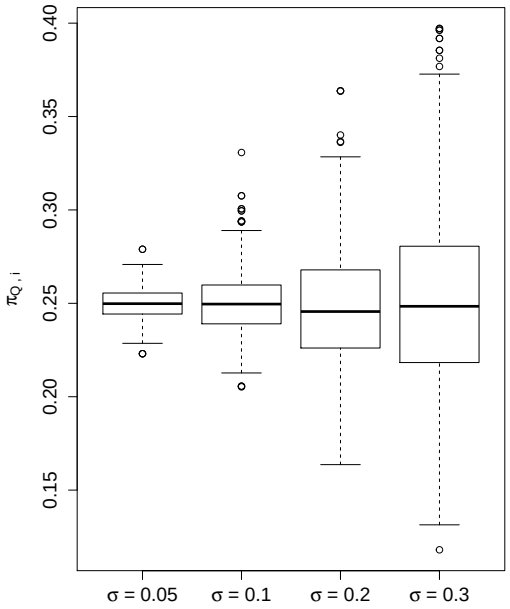


Figure 3.8: Boxplot of the prior for the first element of the stationary distribution for different values of the perturbation standard deviation σ conditional on the rate matrix Q^H(the priors for the rest of the elements of the stationary distribution are the same due to symmetry). Increasing the value of σ clearly increases the spread in the prior for the stationary probabilities.

posterior samples of sizes at least 5K. These samples were obtained by running the algorithm for at least 1000K iterations, discarding about half of the iterations as burn-in and then thinning by taking every 100-th iterate to reduce autocorrelation. Convergence was diagnosed using the procedure described in Section 2.6.3. This involved initialising two MCMC chains at different starting points and graphically comparing the chains through properties based on model parameters and the relative frequencies of sampled clades. In all cases, the graphical diagnostics gave no evidence of any lack of convergence.

Figure B.1 in Appendix B displays the posterior probabilities of the root splits for data sets simulated with five different values of σ and analysed with the Yule prior. The plots show that when $\sigma = 0$ the posterior of the root splits is very similar to the prior, as expected, because when $\sigma = 0$ the data contain no information about the root. As σ increases, the root is often inferred better, with $\sigma = 0.3$ demonstrating the best root inference of all analysed values of σ. However, the analyses of nine simulated data set for each value of σ do not show similar behaviour. There is substantial variability between the data sets, even those simulated with the same rate matrix, and the true root split is not inferred in all cases. The unrooted topologies, however, are inferred with the highest

posterior probabilities for all values of σ (Figure B.2, Appendix B). This suggests that in addition to inferring the unrooted topology, we can use the the NR model to extract some information about the root. Moreover, as expected, the greater the degree of non-reversibility, the stronger the signal from the data. The results of the analysis of the same data sets performed with the structured uniform prior are similar to the Yule prior (Figures C.1 and C.2, Appendix C).

Block Two

In a Bayesian analysis, the posterior distribution reflects information from both the prior and the data. When the prior and likelihood are comparably concentrated, but in conflict, the posterior can only represent a middle ground. In phylogenetics, inferences are known to be highly sensitive to the choice of prior for branch lengths and the topology itself (Yang & Rannala, 2005; Alfaro & Holder, 2006). Motivated by the kinds of conflicts that are likely to arise in the analysis of real biological data, we consider the robustness of posterior root inferences to conflicting prior and likelihood information concerning the rooted topology and branch lengths. An Exp(10) prior for branch lengths which we adopt in our analyses, asserts a strong prior belief that edges will be reasonably short. Therefore, given an unrooted topology which contains a long branch, the prior will typically support placement of the root midway along this branch in order to break it up into two shorter ones. The Yule prior for rooted topologies assigns a (near) uniform distribution to all root types. However, there are generally many fewer trees of unbalanced types, like $1 : n-1$, than there are of more balanced types like $n/2 : n/2$ for n even or $(n-1)/2 : (n+1)/2$ for n odd. It follows that a topology which is more balanced will typically receive more prior mass than a topology which is more unbalanced. In the remainder of this subsection we therefore use simulation to examine posterior robustness in cases where prior-likelihood conflict arises due to a data generating tree which is unbalanced or which contains a long branch.

We base our simulations on an unrooted 30-taxon tree derived from a recent analysis (Figure 3.9) (Williams *et al.*, 2012). This tree describes the relationships between the Archaea and the eukaryotes. We investigate the support for two competing hypotheses about the tree of life (Section 1.2): (i) the three-domains hypothesis, according to which the root of the tree comprising the Archaea and the eukaryotes is placed on the branch separating between the monophyletic Archaea and the eukaryotes (branch E_1 with the length of 1.3), and (ii) the eocyte hypothesis which places the root within the paraphyletic Archaea, (branch E_2 with the length of 0.1). Based on this unrooted tree, we construct six different rooted trees by changing the placement of the root and the length of branch E_1 according to Table 3.1.

Trees 1, 3 and 5 are fairly balanced with root type $11 : 19$, whilst Trees 2, 4 and 6 are

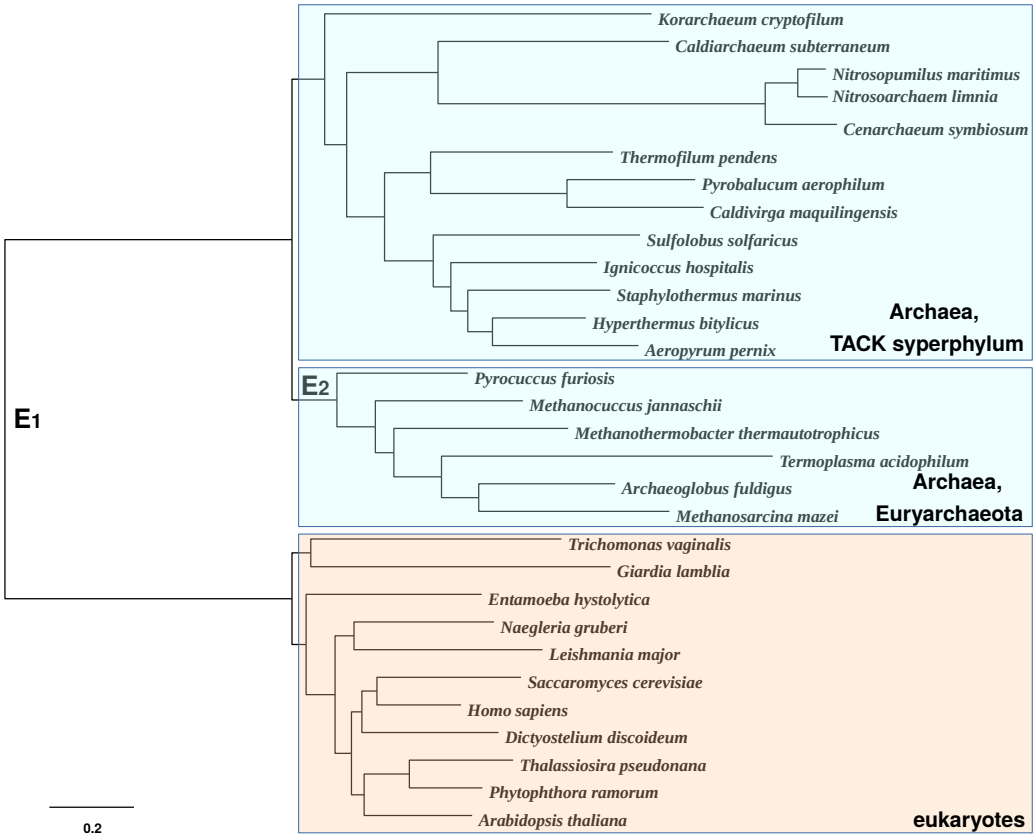


Figure 3.9: An unrooted 30-taxon tree derived from a recent analysis (Williams *et al.*, 2012). The root on branch E_1 corresponds to the three-domains hypothesis (located between the monophyletic Archaea and the eukaryotes), while the root on branch E_2 corresponds to the eocyte hypothesis (located within the paraphyletic Archaea separating the Euryarchaeota from the clade comprising the TACK superphylum and the eukaryotes).

Table 3.1: Six rooted trees for the block two of the simulations. The trees have an unrooted topology of the tree shown in Figure 3.9 but differ in the placement of the root and the length of edge E_1.

Tree	Root edge	Length of E_1
1	E_1	1.3
2	E_2	1.3
3	E_1	0.1
4	E_2	0.1
5	E_1	0.3
6	E_2	0.3

more unbalanced with root type 6 : 24. The Yule prior assigns almost 30% more mass to the former rooted topology. In Trees 1 and 2 and, to a lesser extent, Trees 5 and 6, the unrooted topology contains a long internal branch. In Trees 3 and 4 this internal branch is short. Given the unrooted tree depicted in Figure 1, the prior will therefore support placement of the root on branch E_1 , particularly if this branch is long. We use the NR model to simulate a rate matrix Q with $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)$, $\kappa = 2$ and $\sigma = 0.3$. In turn, this rate matrix is used to simulate three different alignments for each tree. These alignments are then analysed under the NR model with the Yule prior.

(i) Tree 1.

Tree 1 is rooted on the long branch E_1 . Clearly the likelihood for data generated from this tree will support the correct placement of the root. Moreover, for the reasons expressed above, the prior will also support rooting on edge E_1 . It is not surprising, therefore, that we find the posterior is very concentrated around the true root position (Figure 3.10).

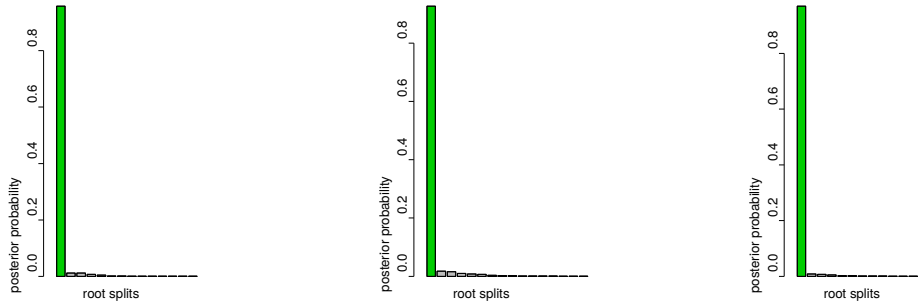


Figure 3.10: Posterior distribution of the root splits for three different alignments simulated under Tree 1. The true root split has high posterior support, possibly because it is heavily favoured by the prior. The green bar here and on all the following plots corresponds to the true root split.

(ii) Tree 2.

In Tree 2, the root is placed on the much shorter branch E_2 , creating a fairly unbalanced unrooted topology with a long interior branch E_1 . As such, data generated under this tree will favour the correct root position on edge E_2 , but the prior will favour a root on branch E_1. This creates prior-likelihood conflict. As expected, we find that the posterior probability of the true root drops substantially in comparison to the analysis for Tree 1 and in two of the three analyses, the posterior offers more support to a root on edge E_1 (Figure 3.11).

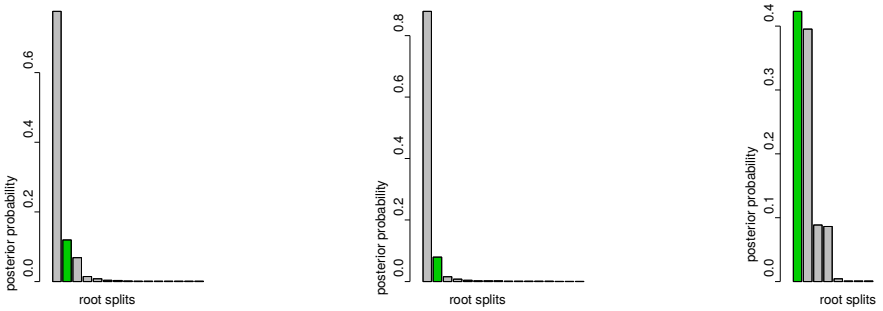


Figure 3.11: Posterior distribution of the root splits for three different alignments simulated under Tree 2. The tree is rooted on a relatively short branch. The support for the true root decreases in comparison to the analysis for Tree 1, presumably because of the presence of the long internal branch.

(iii) Tree 3.

Tree 3 has the same rooted topology as Tree 1 but the root branch E_1 is now much shorter and the unrooted topology does not contain any long edges. As for Tree 1, prior-likelihood conflict does not arise but there is no longer such pronounced prior support for placement of the root on edge E_1 . Nevertheless, we find that the posterior is still concentrated around the true root position (Figure 3.12.)

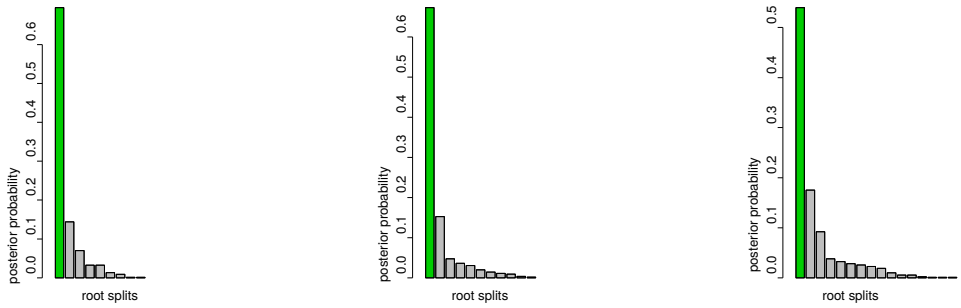


Figure 3.12: Posterior distribution of the root splits for three different alignments simulated under Tree 3. The tree is balanced and has no long branches, so the root is inferred with the highest posterior support.

(iv) Tree 4.

Tree 4 has the same rooted topology as Tree 2 but the long interior branch E_1 is now shortened to 0.1. Although the Yule prior generally favours more balanced trees than Tree 4, the prior for branch lengths no longer offers overwhelming support to placement of the root on edge E_1 . We find that the true root can now be recovered as the posterior mode

(Figure 3.13) but with less support than in the analysis for Tree 3.

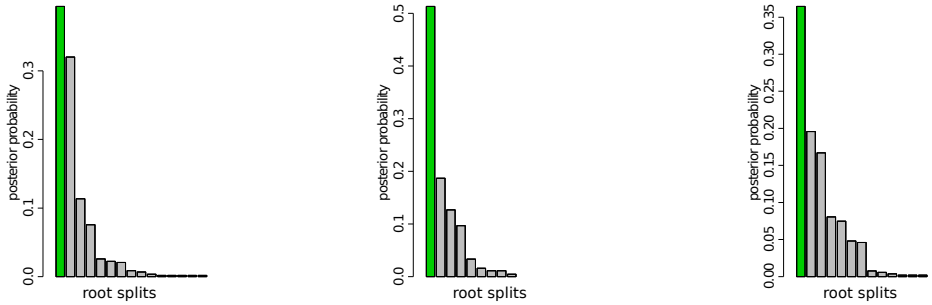


Figure 3.13: Posterior distribution of the root splits for three different alignments simulated under Tree 4. The tree has no long branches but it is less balanced than Tree 3. Still, the root is inferred with the highest posterior support.

(v) Tree 5.

Tree 5 has the same rooted topology as Trees 1 and 3, but the root edge E_1 has length 0.3, which lies between the corresponding values for Trees 1 and 3. As expected, we find that the true root is inferred as the posterior mode (Figure 3.14), and the posterior is less (more) concentrated around the mode in comparison to the analysis of Tree 1 (Tree 3) (Figure 3.14).

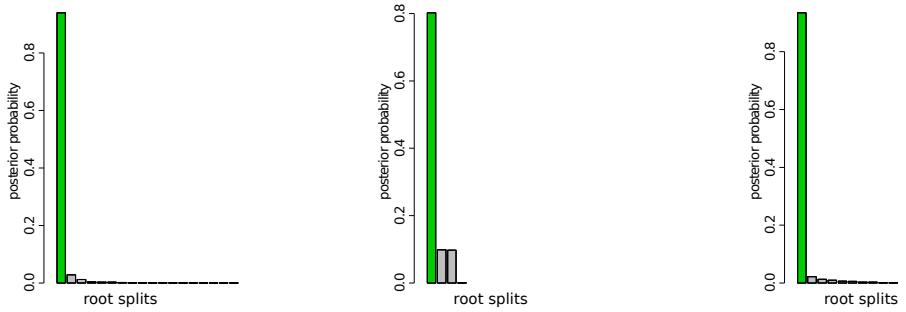


Figure 3.14: Posterior distribution of the root splits for three different alignments simulated under Tree 5. The tree is balanced and the root edge is relatively long, so the true root split is inferred quite high posterior support.

(vi) Tree 6.

Tree 6 has the same rooted topology as Trees 2 and 4, but the internal edge E_1 has length 0.3, which lies between the corresponding values for Trees 2 and 4. The unrooted topology has a moderately long interior edge and the rooted topology is unbalanced, lead-

ing to some prior-likelihood conflict. We find that a root on edge E_1 sometimes receives more posterior support than the true root (Figure 3.15), although, as expected, this effect is less pronounced than in the analysis for Tree 2.

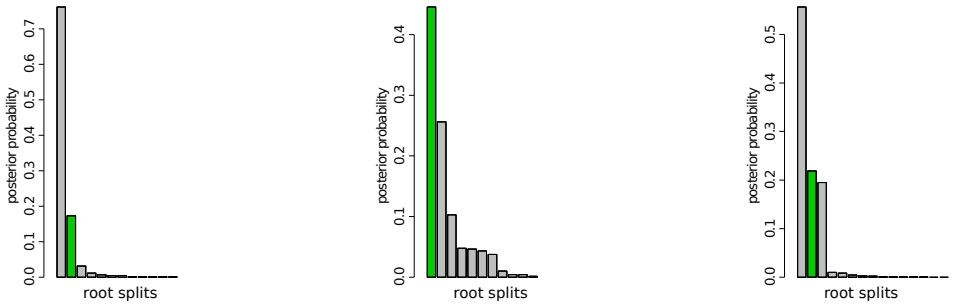


Figure 3.15: Posterior distribution of the root splits for three different alignments simulated under Tree 6. This tree has a relatively long internal branch. The support for the true root split decreases in comparison to the same rooted tree with no long internal branch.

This simulation experiment illustrates the sensitivity of root inferences to conflict between the prior and the likelihood. The effect of a mismatch in information about branch lengths is particularly noticeable. Given a particular unrooted topology, whilst the likelihood might support the presence of a long branch in the corresponding rooted tree, an Exp(10) prior does not (Figure 3.16), and therefore favours placement of the root on the

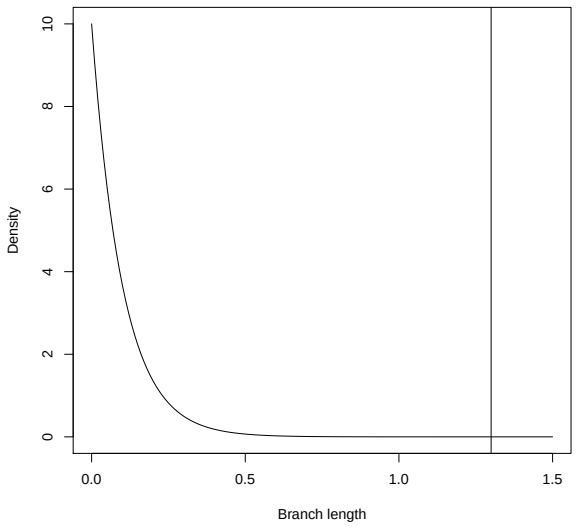


Figure 3.16: Prior distribution of the branch length, Exp(10). Vertical line represents the branch length of 1.3. This plot shows that branches longer than approximately 0.3 have negligible prior support.

long edge. Long branches are not uncommon since overall rates of evolution can differ over the tree (Foster, 2004) and therefore lead to fast lineages represented by the long branches. Ideally constructing a more flexible prior which more explicitly models topology and branch lengths jointly will contribute to better root inference. However, given the absence of very long branches, our results show that the model is still able to extract information from the data about the root even in the face of prior-likelihood conflict.

Run times

The analysis of an alignment with 2000 sites and 30 taxa took approximately 3 days to obtain 500 000 MCMC iterations.

3.2 Two components model

Under the NR model, departures from the HKY85-structure could lead to a non-reversible model or possibly just a more general reversible rate matrix. As such the two types of deviation are confounded and so for any given data set, learning that σ is large does not necessarily provide evidence of non-reversibility (Figure 3.17). The NR2 model addresses this issue, thereby aiding model interpretation, by using a two-stage process to perturb the underlying HKY85 rate matrix Q^H.

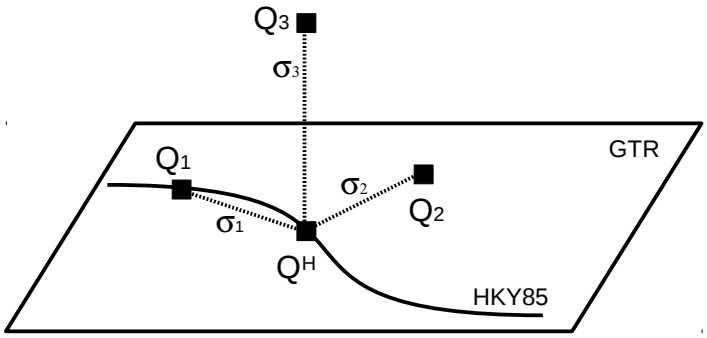


Figure 3.17: A figurative illustration of the space of rate matrices. The curve represents the space of HKY85 matrices, the plane represents the space of GTR matrices which contains HKY85 matrices. The Q^H matrix might be perturbed with σ_1 to obtain another HKY85 matrix Q_1. It might also be perturbed with σ_2 to obtain a general reversible matrix Q_2, or it might be perturbed with σ_3 to obtain a non-reversible matrix Q_3. Hence, large σ does not necessarily provide evidence of non-reversibility.

3.2.1 Model description

The two-stage perturbation process is designed as follows. The first perturbation is within the space of GTR matrices, perpendicular to the subspace of HKY85 matrices, leading to a

reversible rate matrix denoted Q^R. The second perturbation acts on Q^R and is within the space of general rate matrices but perpendicular to the subspace of GTR matrices, leading to a general non-reversible rate matrix denoted Q. These two random perturbations have different variance parameters σ_R^2 and σ_N^2 respectively (Figure 3.18). Orthogonality ensures that (at least locally) Q^R is *not* an HKY85 matrix, and Q is *not* reversible.

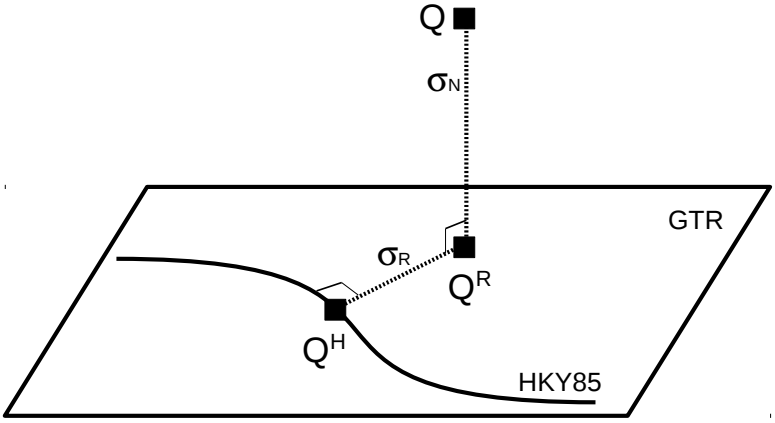


Figure 3.18: Two-stage process to perturb the underlying HKY85 rate matrix Q^H. The perturbation within the set of reversible matrices is performed using σ_R, while the perturbation into the non-reversible part of the rate matrix space is performed with σ_N.

The two-stage perturbation relies upon the underlying geometry of the space of Markov rate matrices, and is achieved in the following way. We work on a log-scale element-wise with all matrices, ignoring diagonal elements. The set of all possible 4-by-4 rate matrices M is therefore identified with $\mathbb{R}^{12}$ which we equip with the standard inner product. The sets of HKY85 matrices and GTR matrices form nested sub-manifolds of M. Recall that working element-wise on a log-scale, the off-diagonal elements of the rate matrix of the NR model can be expressed as, for $i \neq j$

$$\log q_{ij} = \log q_{ij}^H + \epsilon_{ij}, \tag{3.4}$$

where the ϵ_{ij} are independent $N(0,\ \sigma^2)$ quantities. The element-wise log of the HKY85 matrix Q^H in equation (3.4) is

$$\sum_{i=1}^{4} \hat{\pi}_i \mathbf{s}\, \mathbf{e}_i^T + \hat{\kappa}\left(\mathbf{e}_1\mathbf{e}_2^T + \mathbf{e}_2\mathbf{e}_1^T + \mathbf{e}_3\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_3^T\right),$$

where $(\hat{\pi}_1,\ \hat{\pi}_2,\ \hat{\pi}_3,\ \hat{\pi}_4) = (\log \pi_A,\ \log \pi_G,\ \log \pi_C,\ \log \pi_T)$, $\hat{\kappa} = \log \kappa$, $\mathbf{e}_i$ is the i-th standard basis vector of $\mathbb{R}^4$ and $\mathbf{s} = (1,\ 1,\ 1,\ 1)^T$. By differentiating with respect to the parameters $\hat{\pi}_1,\ \hat{\pi}_2,\ \hat{\pi}_3$ and $\hat{\kappa}$ we obtain 4 linearly independent vectors in M which are locally tangent to the sub-manifold of HKY85 matrices at Q^H, and we denote these

$V_1,\ V_2,\ V_3,\ V_4$. (Differentiating with respect to $\hat{\pi}_4$ gives a tangent vector contained in the span of $V_1,\ V_2,\ V_3$.) The tangent vectors in M correspond to the 4-by-4 matrices

$$V_i = \mathbf{s}\,\mathbf{e}_i^T - \exp(\hat{\pi}_i - \hat{\pi}_4)\mathbf{s}\,\mathbf{e}_4^T \quad \text{for } i = 1, 2, 3,$$

and

$$V_4 = \mathbf{e}_1\mathbf{e}_2^T + \mathbf{e}_2\mathbf{e}_1^T + \mathbf{e}_3\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_3^T.$$

The element-wise log of the general GTR matrix is

$$\sum_{i=1}^{4} \hat{\pi}_i \mathbf{s}\,\mathbf{e}_i^T + \sum_{i<j} \hat{\rho_{ij}} \left(\mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T\right),$$

where $\hat{\rho}_{ij}$ is the log of the exchangeability parameter ρ_{ij} (see Section 2.2.5). By differentiating with respect to the $\hat{\rho}_{ij}$ parameters, it is straightforward to obtain tangent vectors $V_5, \ldots, V_9$ to the sub-manifold of GTR matrices at Q^H, such that the set $V_1, \ldots, V_9$ is linearly independent:

$$\begin{aligned}
V_5 &= (\mathbf{e}_1\mathbf{e}_2^T + \mathbf{e}_2\mathbf{e}_1^T) - (\mathbf{e}_3\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_3^T) \\
V_6 &= (\mathbf{e}_1\mathbf{e}_3^T + \mathbf{e}_3\mathbf{e}_1^T) + (\mathbf{e}_2\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_2^T) \\
V_7 &= (\mathbf{e}_1\mathbf{e}_3^T + \mathbf{e}_3\mathbf{e}_1^T) - (\mathbf{e}_2\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_2^T) \\
V_8 &= (\mathbf{e}_1\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_1^T) + (\mathbf{e}_2\mathbf{e}_3^T + \mathbf{e}_3\mathbf{e}_2^T) \\
V_9 &= (\mathbf{e}_1\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_1^T) - (\mathbf{e}_2\mathbf{e}_3^T + \mathbf{e}_3\mathbf{e}_2^T)
\end{aligned}$$

Finally, standard linear algebra can be used to extend this collection to a basis $V_1, \ldots, V_{12}$ of $\mathbb{R}^{12}$:

$$\begin{aligned}
V_{10} &= (\mathbf{e}_1\mathbf{e}_2^T - \mathbf{e}_2\mathbf{e}_1^T) - (\mathbf{e}_3\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_3^T) \\
V_{11} &= (\mathbf{e}_1\mathbf{e}_3^T - \mathbf{e}_3\mathbf{e}_1^T) - (\mathbf{e}_2\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_2^T) \\
V_{12} &= (\mathbf{e}_1\mathbf{e}_4^T - \mathbf{e}_4\mathbf{e}_1^T) + (\mathbf{e}_2\mathbf{e}_3^T - \mathbf{e}_3\mathbf{e}_2^T)
\end{aligned}$$

Thus, vectors $V_1, \ldots, V_{12}$ form a 12-by-12 matrix with the following columns:

$$(0,\ 0,\ -\exp(\hat{\pi}_1 - \hat{\pi}_4),\ 1,\ 0,\ -\exp(\hat{\pi}_1 - \hat{\pi}_4),\ 1,\ 0,\ -\exp(\hat{\pi}_1 - \hat{\pi}_4),\ 1,\ 0,\ 0)^T$$
$$(1,\ 0,\ -\exp(\hat{\pi}_2 - \hat{\pi}_4),\ 0,\ 0,\ -\exp(\hat{\pi}_2 - \hat{\pi}_4),\ 0,\ 1,\ -\exp(\hat{\pi}_2 - \hat{\pi}_4),\ 0,\ 1,\ 0)^T$$
$$(0,\ 1,\ -\exp(\hat{\pi}_3 - \hat{\pi}_4),\ 0,\ 1,\ -\exp(\hat{\pi}_3 - \hat{\pi}_4),\ 0,\ 0,\ -\exp(\hat{\pi}_3 - \hat{\pi}_4),\ 0,\ 0,\ 1)^T$$
$$(1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1)^T$$
$$(1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ ,-1,\ 0,\ 0,\ -1)^T$$
$$(0,\ 1,\ 0,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 0,\ 1,\ 0)^T$$
$$(0,\ 1,\ 0,\ 0,\ 0,\ -1,\ 1,\ 0,\ 0,\ 0,\ -1,\ 0)^T$$
$$(0,\ 0,\ 1,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 1,\ 0,\ 0)^T$$
$$(0,\ 0,\ 1,\ 0,\ -1,\ 0,\ 0,\ -1,\ 0,\ 1,\ 0,\ 0)^T$$
$$(1,\ 0,\ 0,\ -1,\ 0,\ 0,\ 0,\ 0,\ -1,\ 0,\ 0,\ 1)^T$$
$$(0,\ 1,\ 0,\ 0,\ 0,\ -1,\ -1,\ 0,\ 0,\ 0,\ 1,\ 0)^T$$
$$(0,\ 0,\ 1,\ 0,\ 1,\ 0,\ 0,\ -1,\ 0,\ -1,\ 0,\ 0)^T$$

Next, the QR factorisation algorithm is applied to this matrix to obtain an orthonormal basis of tangent vectors $W_1, \ldots, W_{12}$ which is used to perturb Q^H. First, Q^H is perturbed using $\nu_1, \ldots, \nu_5$ to obtain a GTR matrix Q^R where, for $i \neq j$

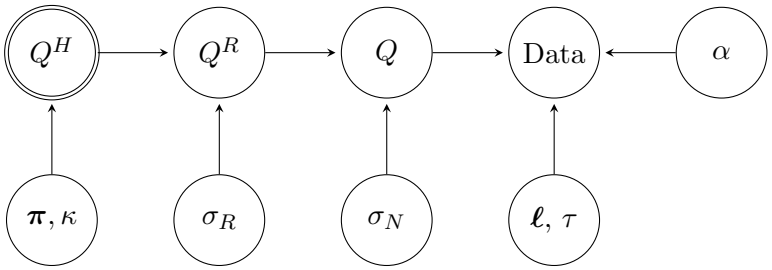
$$\log q_{ij}^R = \log q_{ij}^H + \sum_{k=5}^{9} \nu_{k-4} W_{kij},$$

and the ν_k are independent $N(0, \sigma_R^2)$ and W_{kij} is the (i,j)-th element of the 4-by4 matrix W_k. The choice of basis $W_1, \ldots, W_{12}$ ensures that this perturbation is locally orthogonal to the sub-manifold of HKY85 matrices. The second stage perturbs Q^R into the space of non-reversible rate matrices using η_1, η_2, η_3: for $i \neq j$

$$\log q_{ij} = \log q_{ij}^R + \sum_{k=10}^{12} \eta_{k-9} W_{kij},$$

and the η_k are independent $N(0, \sigma_N^2)$ quantities. This perturbation is locally perpendicular to the sub-manifold of GTR matrices in M. The equation determines the off-diagonal elements of the non-reversible rate matrix Q, while the diagonal elements are fixed in order to make the row sums zero. The size of the perturbation variance σ_R^2 can be thought of as representing the extent to which the rate matrix Q departs from the class of HKY85 rate matrices remaining within the class of reversible models, while σ_N^2 represents the extent to which Q departs from being reversible. The hierarchical model for sequence data has

the following structure:



3.2.2 Prior

The parameters of the NR2 model are: the composition vector $\boldsymbol{\pi}$, the transition-transversion rate ratio κ, the perturbation standard deviation within the set of reversible matrices σ_R, the perturbation standard deviation within the set of non-reversible matrices σ_N, the gamma shape heterogeneity parameter α, the branch lengths $\boldsymbol{\ell}$ and the rooted topology τ. We also have latent variables comprising $\nu_1, \ldots, \nu_5$ for the reversible perturbation, and $\eta_1,\ \eta_2,\ \eta_3$ for the non-reversible perturbation. The prior distribution of these unknowns is given by

$$\pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau) = \pi(\boldsymbol{\pi})\pi(\kappa)\pi(\sigma_R)\pi(\sigma_N)\pi(\boldsymbol{\nu}|\sigma_R)\pi(\boldsymbol{\eta}|\sigma_N)\pi(\alpha)\pi(\boldsymbol{\ell})\pi(\tau).$$

The rate heterogeneity parameter α, branch lengths $\boldsymbol{\ell}$, rooted topology τ and the parameters $\boldsymbol{\pi}$ and κ of the reversible Q^H matrix are assigned the same priors as those used for the NR model. Both perturbation standard deviations are assigned the same prior as their analogue, σ, in the NR model, i.e. $\sigma_R \sim \text{Exp}(2.3)$ and $\sigma_N \sim \text{Exp}(2.3)$. As discussed in Section 3.2.1, $\nu_i \sim N(0, \sigma_R^2)$ for $i = 1, \ldots, 5$ independently, and $\eta_i \sim N(0, \sigma_N^2)$ for $i = 1,\ 2,\ 3$ independently.

3.2.3 Posterior inference via MCMC

Here the posterior distribution of the unknowns is given by

$$\begin{aligned}\pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau|D) \propto \pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \alpha, \boldsymbol{\ell}, \tau) \times \pi(\boldsymbol{\nu}|\sigma_R) \times \pi(\boldsymbol{\eta}|\sigma_N) \\ \times \pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)\end{aligned}$$

and an analogous Metropolis-within-Gibbs algorithm is used to generate posterior samples:

(i) Metropolis-Hastings step for the composition vector $\boldsymbol{\pi}$:

Prior: $\boldsymbol{\pi} \sim \mathcal{D}(\alpha_\pi \boldsymbol{\pi}_0),\ \alpha_\pi = 4,\ \boldsymbol{\pi}_0 = (0.25,\ 0.25,\ 0.25,\ 0.25)$.

Proposal: $\boldsymbol{\pi}' \sim \mathcal{D}(a_\pi \boldsymbol{\pi})$, where $\boldsymbol{\pi}$ is the current value, a_π is a tuning parameter.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\pi}')}{\pi(\boldsymbol{\pi})} \times \frac{q(\boldsymbol{\pi}|\boldsymbol{\pi}')}{q(\boldsymbol{\pi}'|\boldsymbol{\pi})} \times \frac{\pi(D|\boldsymbol{\pi}', \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \prod_{i=1}^{4} \frac{\Gamma(a_\pi \pi_i)}{\Gamma(a_\pi \pi_i')} \pi_i^{(a_\pi \pi_i' - \alpha_\pi \pi_{0i})} \pi_i'^{(\alpha_\pi \pi_{0i} - a_\pi \pi_i)} \times \frac{\pi(D|\boldsymbol{\pi}', \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

(ii) Metropolis-Hastings step for the transition-transversion rate ratio κ:

Prior: $\kappa \sim \text{LN}(\log \kappa_0, \xi^2)$, $\kappa_0 = 1,\ \xi = 0.8$.
Proposal: $\kappa' \sim \text{LN}(\log \kappa, a_\kappa^2)$, where κ is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\kappa')}{\pi(\kappa)} \times \frac{q(\kappa|\kappa')}{q(\kappa'|\kappa)} \times \frac{\pi(D|\boldsymbol{\pi}, \kappa', \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \exp\left[\frac{1}{2\xi^2}\left\{(\log \kappa)^2 - (\log \kappa')^2 + 2 \log \kappa_0 (\log \kappa' - \log \kappa)\right\}\right] \times \frac{\pi(D|\boldsymbol{\pi}, \kappa', \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

(iii) Metropolis-Hastings step for the reversible perturbation standard deviation σ_R:

Prior: $\sigma_R \sim \text{Exp}(\gamma),\ \gamma = 2.3$.
Proposal: $\sigma_R' \sim \text{LN}(\log \sigma_R,\, a_{\sigma_R}^2)$, where σ_R is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\sigma_R')}{\pi(\sigma_R)} \times \frac{q(\sigma_R|\sigma_R')}{q(\sigma_R'|\sigma_R)} \times \frac{\pi(\boldsymbol{\nu}|\sigma_R')}{\pi(\boldsymbol{\nu}|\sigma_R)} \times \frac{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \frac{\pi(\sigma_R')}{\pi(\sigma_R)} \times \frac{q(\sigma_R|\sigma_R')}{q(\sigma_R'|\sigma_R)} \times \frac{\pi(\boldsymbol{\nu}|\sigma_R')}{\pi(\boldsymbol{\nu}|\sigma_R)} \\ &= \left(\frac{\sigma_R}{\sigma_R'}\right)^4 \exp\left\{\gamma(\sigma_R - \sigma_R') - \frac{1}{2}\left(\frac{1}{\sigma_R'^2} - \frac{1}{\sigma_R^2}\right) \sum_{i=1}^{5} \nu_i^2\right\}. \end{aligned}$$

(iv) Metropolis-Hastings step for the reversible perturbation component $\boldsymbol{\nu}$ (a sweep through all the elements of $\boldsymbol{\nu}$):

Prior: $\nu_k \sim N(0, \sigma_R^2), k = 1,\ \ldots,\ 5$.

Proposal: $\nu_k' \sim N(\nu_k,\, a^2_{\nu_k})$, where ν_k is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\nu}')}{\pi(\boldsymbol{\nu})} \times \frac{q(\nu_k|\nu_k')}{q(\nu_k'|\nu_k)} \times \frac{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}', \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \exp\left\{\sum_{k=1}^{5} \left(\nu_k^2 - \nu_k'^2\right) / \left(2\sigma_R^2\right)\right\} \times \frac{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}', \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

(v) Metropolis-Hastings step for the non-reversible perturbation standard deviation σ_N:

Prior: $\sigma_N \sim \text{Exp}(\gamma),\ \gamma = 2.3$.
Proposal: $\sigma_N' \sim \text{LN}(\log \sigma_N,\, a^2_{\sigma_N})$, where σ_N is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\sigma_N')}{\pi(\sigma_N)} \times \frac{q(\sigma_N|\sigma_N')}{q(\sigma_N'|\sigma_N)} \times \frac{\pi(\boldsymbol{\eta}|\sigma_N')}{\pi(\boldsymbol{\eta}|\sigma_N)} \times \frac{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \frac{\pi(\sigma_N')}{\pi(\sigma_N)} \times \frac{q(\sigma_N|\sigma_N')}{q(\sigma_N'|\sigma_N)} \times \frac{\pi(\boldsymbol{\eta}|\sigma_n')}{\pi(\boldsymbol{\eta}|\sigma_N)} \\ &= \left(\frac{\sigma_N}{\sigma_N'}\right)^2 \exp\left\{\gamma(\sigma_N - \sigma_N') - \frac{1}{2}\left(\frac{1}{\sigma_N'^2} - \frac{1}{\sigma_N^2}\right)\sum_{i=1}^{3} \eta_i^2\right\}. \end{aligned}$$

(vii) Metropolis-Hastings step for the non-reversible perturbation component $\boldsymbol{\eta}$ (a sweep through all the elements of $\boldsymbol{\eta}$):

Prior: $\eta_k \sim N(0, \sigma_N^2),\ k = 1,\ 2,\ 3$.
Proposal: $\eta_k' \sim N(\eta_k,\, a^2_{\eta_k})$, where η_k is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\eta}')}{\pi(\boldsymbol{\eta})} \times \frac{q(\eta_k|\eta_k')}{q(\eta_k'|\eta_k)} \times \frac{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}', \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \exp\left\{\sum_{k=1}^{3} \left(\eta_k^2 - \eta_k'^2\right) / \left(2\sigma_N^2\right)\right\} \times \frac{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}', \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \kappa, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

Metropolis-Hastings steps for the gamma shape heterogeneity parameter α, the branch lengths $\boldsymbol{\ell}$ and the topology τ are the same as those used for the NR model.

3.2.4 Simulation study

The simulations are performed in a similar manner as for the first block of the simulations for the NR model. We use the same rooted tree (Figure 3.6) to create nine alignments for each one of five values of $\sigma_N = 0,\ 0.1,\ 0.25,\ 0.5,\ 1.0$. In all the simulations we used the same value for the reversible perturbation, $\sigma_R = 0.1$. Note, that the case of $\sigma_N = 0$ corresponds to the GTR model. The values of $\sigma_N = 0.1,\ 0.25,\ 0.5,\ 1.0$ were chosen so that in the prior for the stationary distribution, some nucleotides are not heavily favoured over the others (Figure 3.19). We note, that this type of perturbation allows us to use larger values of σ_N in comparison to the values of σ in the NR model, while still maintaining a realistic stationary distribution. As for the NR model, for each value of σ_N the first five alignments were simulated from different rate matrices (data sets 1 - 5), while the last five alignments were simulated from the same rate matrix (data sets 5 - 9). All of the alignments were simulated using a gamma shape heterogeneity parameter simulated from $\text{Ga}(10, 10)$.

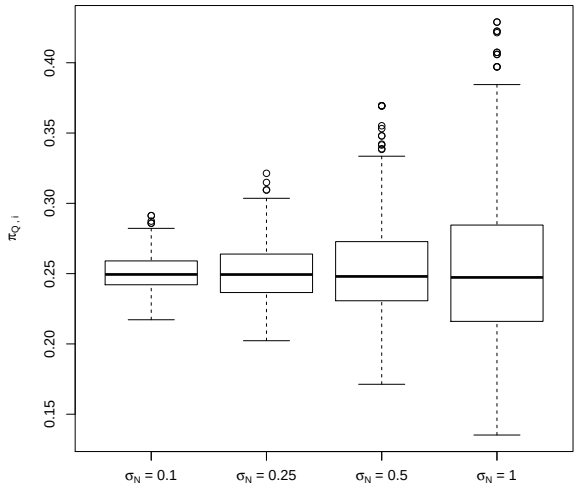


Figure 3.19: Boxplot of the prior for the first element of the stationary distribution for different values of σ_N with $\sigma_R = 0.1$, conditional on the rate matrix Q^H (the priors for the rest of the elements of the stationary distribution are the same due to symmetry). Increasing the value of σ_N clearly increases the spread in the prior for the stationary probabilities.

Figure D.1 in Appendix D displays the posterior probabilities of the root splits for data sets simulated with five different values of σ_N and analysed with the Yule prior. As with the NR model, when $\sigma_N = 0$ the posterior of the root splits is very similar to the prior, as expected given that the data contain no information about the root. As σ_N increases, the root is inferred better, with $\sigma_N = 1$ demonstrating the best root inference of all the values

of σ_N analysed. It is clear, that the NR2 model infers the true root split better than the NR model. For large values of σ_N the true root split has very high posterior support for all cases. This can be explained by the fact that the structure of the NR2 model allows using larger values of the non-reversible perturbation component which is the source of the root information. Indeed, when fitting the NR model to the data simulated under the NR2 model, we obtained very similar root inferences to those summarised in Figure D.1 in Appendix D, with strong posterior support for the correct root position for large σ_N (not shown). The unrooted topologies are inferred with the highest posterior probabilities for all the values of σ (Figure D.2, Appendix D). The analysis of the same data sets performed with the structured uniform prior shows similar results (Figures E.1 and E.2, Appendix E).

Run times

The analysis of an alignment with 2000 sites and 30 taxa took approximately 3 days to obtain 500 000 MCMC iterations (similarly to the NR model).

3.3 Model for Dayhoff re-coding

3.3.1 Model description

The model is based on a GTR model for Dayhoff-recoding which is specified by the following rate matrix:

$$Q^G = (q_{ij}) = \begin{pmatrix} \star & \rho_{12}\pi_2 & \rho_{13}\pi_3 & \rho_{14}\pi_4 & \rho_{15}\pi_5 & \rho_{16}\pi_6 \\ \rho_{12}\pi_1 & \star & \rho_{23}\pi_3 & \rho_{24}\pi_4 & \rho_{25}\pi_5 & \rho_{26}\pi_6 \\ \rho_{13}\pi_1 & \rho_{23}\pi_2 & \star & \rho_{34}\pi_4 & \rho_{35}\pi_5 & \rho_{36}\pi_6 \\ \rho_{14}\pi_1 & \rho_{24}\pi_2 & \rho_{34}\pi_3 & \star & \rho_{45}\pi_5 & \rho_{46}\pi_6 \\ \rho_{15}\pi_1 & \rho_{25}\pi_2 & \rho_{35}\pi_3 & \rho_{45}\pi_4 & \star & \rho_{56}\pi_6 \\ \rho_{16}\pi_1 & \rho_{26}\pi_2 & \rho_{36}\pi_3 & \rho_{46}\pi_4 & \rho_{56}\pi_6 & \star \end{pmatrix},$$

where $\boldsymbol{\pi} = (\pi_1,\ \pi_2,\ \pi_3,\ \pi_4,\ \pi_5,\ \pi_6)$ is the amino-acid frequency vector for six Dayhoff groups and $\rho_{ij},\ i = 1, \ldots, 5,\ j = i+1, \ldots, 6$ are the exchangeability parameters. As with the NR model for DNA, the non-reversibility of this model is achieved by a log-normal perturbation of the off-diagonal elements of the rate matrix Q^G.

3.3.2 Prior

The parameters of the model are: the composition vector $\boldsymbol{\pi}$, the exchangeability parameters $\rho_{ij},\ i = 1, \ldots, 4,\ j = i+1, \ldots, 6$ (ρ_{56} is fixed to 1 for identifiability), the perturbation standard deviation σ, the off-diagonal elements of the rate matrix Q, the shape parameter

for the gamma distribution for the across site variation α, the branch lengths $\boldsymbol{\ell}$ and the rooted topology τ. The prior distribution of the parameters is given by:

$$\pi(\boldsymbol{\pi}, \boldsymbol{\rho}, \sigma, Q, \alpha, \boldsymbol{\ell}, \tau) = \pi(\boldsymbol{\pi})\pi(\boldsymbol{\rho})\pi(\sigma)\pi(Q|\boldsymbol{\pi}, \boldsymbol{\rho}, \sigma)\pi(\alpha)\pi(\boldsymbol{\ell})\pi(\tau).$$

The composition vector $\boldsymbol{\pi}$ is defined on the six-dimensional simplex, that is it has five free positive elements and the sixth positive one is fixed such that the elements of $\boldsymbol{\pi}$ sum to 1. We choose to assign it a Dirichlet prior, that is $\boldsymbol{\pi} \sim \mathcal{D}(\alpha_\pi \boldsymbol{\pi}_0)$, where $\boldsymbol{\pi}_0 = (1/6,\ 1/6,\ 1/6,\ 1/6,\ 1/6,\ 1/6)$ and α_π is a concentration parameter equal to 6. Elements of $\boldsymbol{\rho}$ are assigned a log-normal prior with median 1 and standard deviation 0.9, i.e. $\rho \sim \text{LN}(\log 1, 0.9^2)$. Priors for σ, α, $\boldsymbol{\ell}$ and τ are the same as in the NR model for DNA.

3.3.3 Posterior inference via MCMC

The posterior distribution of the parameters is given by

$$\pi(\boldsymbol{\pi}, \boldsymbol{\rho}, \sigma, Q, \alpha, \boldsymbol{\ell}, \tau|D) \propto \pi(Q|\boldsymbol{\pi}, \boldsymbol{\rho}, \sigma) \times \pi(\boldsymbol{\pi}, \boldsymbol{\rho}, \sigma, \alpha, \boldsymbol{\ell}, \tau) \times \pi(D|Q, \alpha, \boldsymbol{\ell}, \tau).$$

(i) Metropolis-Hastings step for the composition vector $\boldsymbol{\pi}$, the standard deviation σ, the off-diagonal elements of the rate matrix Q, the gamma shape heterogeneity parameter α, and also for the branch lengths $\boldsymbol{\ell}$ and the topology τ are similar as in the NR model for DNA.

(ii) Metropolis-Hastings step for the exchangeability parameters $\rho_{ij},\ i = 1, \ldots, 4,\ j = i+1, \ldots, 6$ (a sweep through all ρ_{ij}):

Prior: $\rho_{ij} \sim \text{LN}(\log \rho_0, \lambda^2),\ \rho_0 = 1,\ \lambda = 0.9$.
Proposal: $\rho'_{ij} \sim \text{LN}(\log \rho_{ij},\, a_\rho^2)$, where ρ_{ij} is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned}
A &= \frac{\pi(\rho'_{ij})}{\pi(\rho_{ij})} \times \frac{q(\rho_{ij}|\rho'_{ij})}{q(\rho'_{ij}|\rho_{ij})} \times \frac{\pi(Q|\boldsymbol{\pi}, \boldsymbol{\rho}', \sigma)}{\pi(Q|\boldsymbol{\pi}, \boldsymbol{\rho}, \sigma)} \times \frac{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \alpha, \boldsymbol{\ell}, \tau)} \\
&= \frac{\pi(\rho'_{ij})}{\pi(\rho_{ij})} \times \frac{q(\rho_{ij}|\rho'_{ij})}{q(\rho'_{ij}|\rho_{ij})} \times \frac{\pi(Q|\boldsymbol{\pi}, \boldsymbol{\rho}', \sigma)}{\pi(Q|\boldsymbol{\pi}, \boldsymbol{\rho}, \sigma)} \\
&= \exp\left[\frac{1}{2\lambda^2}\left\{(\log \rho_{ij})^2 - (\log \rho'_{ij})^2 + 2\log \rho_0(\log \rho'_{ij} - \log \rho_{ij})\right\}\right] \\
&\qquad \times \exp\left[\frac{1}{2\sigma^2}\sum_{i \neq j}\left\{(\log q_{ij}^G)^2 - (\log q_{ij}^{G'})^2 + 2\log q_{ij}(\log q_{ij}^{G'} - \log q_{ij}^G)\right\}\right],
\end{aligned}$$

and $q_{ij}^{G'}$ are the off-diagonal elements of the GTR rate matrix computed with ρ'_{ij}.

3.3.4 Simulation study

In order to simulate the alignments, we first fixed the underlying reversible GTR rate matrix using the following values of $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$:

$$\begin{aligned}\boldsymbol{\rho} = (&0.9187689,\ 0.563163,\ 0.5723444,\ 0.1938824,\ 1.412338,\\ &1.849246,\ 0.2421764,\ 0.1396191,\ 0.1699077,\ 0.2715765,\\ &0.8613947,\ 0.39612,\ 0.7532268,\ 0.9419012,\ 0.9803037),\\ \boldsymbol{\pi} = (&0.25,\ 0.2,\ 0.15,\ 0.2,\ 0.15,\ 0.05).\end{aligned}$$

The values of $\boldsymbol{\rho}$ were calculated based on the exchangeability values of the empirical LG model (averaged over the Dayhoff categories). The values of $\boldsymbol{\pi}$ are proportional to the numbers of the amino-acids in each category. We then apply a log-normal perturbation on the off-diagonal elements of the rate matrix Q^G using the perturbation standard deviation $\sigma = 0.3$.

Here we compare the inference for two different root placements on the unrooted tree from Williams *et al.* (2012) (Figure 3.9). Figure 3.20 shows the posterior distribution of the root splits and the unrooted topologies for the data simulated under the tree rooted according to the three-domains hypothesis (on the longest edge E_1), while Figure 3.21 shows the posterior probabilities of the root splits and the unrooted topologies for the data simulated under the tree rooted according to the eocyte hypothesis (on edge E_2). While the unrooted topology is inferred well for both trees, clearly rooting is easier when the data are simulated under the more balanced tree rooted on the longest edge.

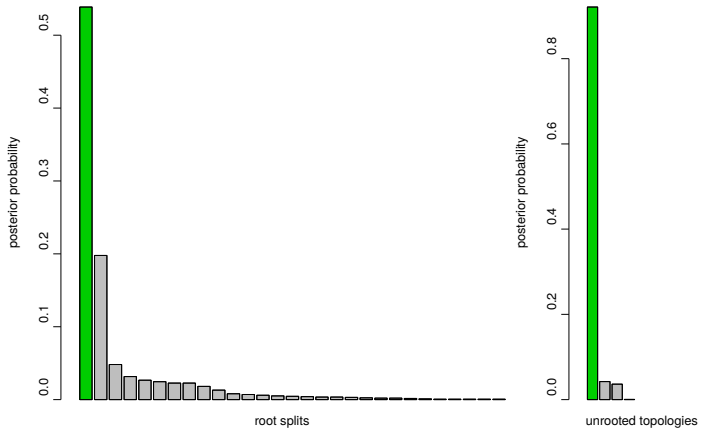


Figure 3.20: Posterior probabilities of the root splits (left) and the unrooted topologies (right) for the non-reversible model for Dayhoff-recoding. The alignment was simulated under the tree shown in Figure 3.9, rooted according to the three-domains hypothesis (root on branch E_1). The true root split and the true unrooted topology are recovered as the posterior mode.

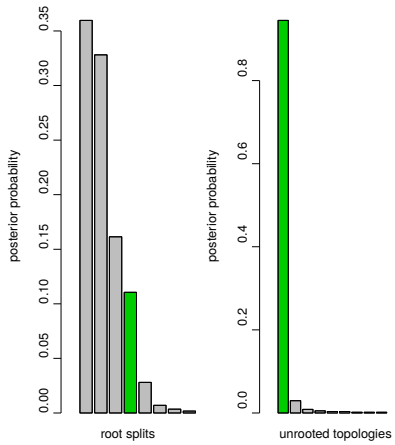


Figure 3.21: Posterior probabilities of the root splits (left) and the unrooted topologies (right) for the non-reversible model for Dayhoff-recoding. The alignment was simulated under the tree shown in Figure 3.9, rooted according to the eocyte hypothesis (root on branch E_2). While the unrooted topology has very high posterior support, the posterior probability of the true root decreases in comparison to the analysis for the three-domains tree shown in Figure 3.20, presumably due to the presence of the long internal branch.

Run times

The analysis of an alignment with 2000 sites and 30 taxa took approximately 6 days to obtain 250 000 MCMC iterations.

Summary

In this chapter we have presented two hierarchical non-reversible models which are centered on a reversible rate matrix, and which also differ in the structure of the perturbation. The first model uses one perturbation component which allows a departure from the HKY85 structure. In contrast, the second model utilises two perturbation components: one which allows a departure from the HKY85 structure towards the general time-reversible structure, and another which allows a departure from the general time-reversible structure. For each model we performed a simulation study for the data with different values of the perturbation parameter. The study showed that the larger the level of non-reversibility in the data the better the root inference is for both models, suggesting that a large degree of non-reversibility provides a better signal for the position of the root. We have also investigated the sensitivity of the model to the conflict between the prior and the likelihood concerning the rooted topology and branch lengths. The simulation study showed that the model is quite robust to the prior-likelihood conflict, given the absence of very long branches.

Chapter 4

Non-stationary substitution models

This chapter focuses on two non-stationary models. The first model is based on the reversible HKY85 model but incorporates non-stationarity by allowing the evolutionary process to start from a distribution that can differ from the equilibrium distribution. The second model adopts the same idea of modelling non-stationarity, but it is based on the non-reversible NR model described in Chapter 3.

4.1 Non-stationary reversible model

The models described in Chapter 3 are stationary. This means that the probability of being in each state (e.g. each nucleotide for DNA) does not change over time and the probabilities of transitioning between states over some time interval depend only on the size of that interval and not on its position in time (meaning that the base composition is constant over time). This assumption is unrealistic from a biological point of view. If it was true, the average composition of the nucleotides would have remained unchanged throughout evolutionary time. Thus stationarity assumption imply all taxa in alignment had sequences with approximately the same proportions of the four nucleotides. Many data sets show evidence that this is not the case (Yang & Roberts, 1995; Foster, 2004; Cox *et al.*, 2008). In fact, the composition is known to change over time, for example due to adaptation to different environmental conditions (Penny *et al.*, 2001; Lopez *et al.*, 2002). It has been shown that failing to accommodate such an important feature of the evolutionary process is one of the reasons for a failure to obtain the correct topology (Foster, 2004). Here, we accommodate non-stationarity by introducing an additional composition vector $\boldsymbol{\pi}_{root}$ which specifies a nucleotide frequency at the root vertex. Thus, the model has two composition vectors: the composition at the root $\boldsymbol{\pi}_{root}$ and the stationary com-

position $\boldsymbol{\pi}$. In other words, the evolutionary process is not assumed to have started at the equilibrium distribution. We use an HKY85 rate matrix model to describe the substitution of nucleotides along branches. Even though the underlying model is reversible, the whole process is non-stationary: the theoretical stationary distribution derived from the rate matrix by $\boldsymbol{\pi}Q = \mathbf{0}$ can be different from the composition at the root vertex. The parameters of the model are: the composition vector at the root vertex $\boldsymbol{\pi}_{root}$, the theoretical stationary distribution $\boldsymbol{\pi}$, the transition-transversion rate ratio κ, the gamma shape heterogeneity parameter α, the branch lengths $\boldsymbol{\ell}$ and the rooted topology τ.

A non-stationary model in which the initial distribution is not the equilibrium distribution has been considered previously in a maximum likelihood framework (Yap & Speed, 2005). It has been found to fit the data better than the stationary non-reversible model. In that study, however, the topology was fixed and the number of taxa was small (no more than nine). Here, we do not fix the unrooted topology and extend the inferential algorithm to allow inference of rooted trees with a greater number of taxa.

4.1.1 Prior

We assume that $\boldsymbol{\pi}$, κ, α, $\boldsymbol{\ell}$ and τ are independent a priori and assign the same prior as in the NR model, i.e. a Dirichlet prior for the composition ($\boldsymbol{\pi} \sim \mathcal{D}(1,\ 1,\ 1,\ 1)$), a log-normal prior for the transition-transversion rate ratio ($\kappa \sim \mathrm{LN}(0, 0.8^2)$) and a gamma prior for the gamma shape heterogeneity parameter ($\alpha \sim \mathrm{Ga}(10, 10)$). However, the distribution at the root is not assumed to be independent of the rate matrix. Rather, its prior is centred on the theoretical stationary distribution $\boldsymbol{\pi}$: $\boldsymbol{\pi}_{root}|\boldsymbol{\pi} \sim \mathcal{D}(k\boldsymbol{\pi})$, where k is a concentration parameter. The prior distribution of the parameters is given by

$$\pi(\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau) = \pi(\boldsymbol{\pi})\pi(\boldsymbol{\pi}_{root}|\boldsymbol{\pi})\pi(\kappa)\pi(\alpha)\pi(\boldsymbol{\ell})\pi(\tau).$$

We explore two possibilities: (a) fixing the concentration parameter k of the Dirichlet prior for the composition at the root vertex $\boldsymbol{\pi}_{root}$; (b) inferring the concentration parameter k.

Fixing the concentration parameter

In order to choose the value of the concentration parameter k, we begin by analysing plots of the marginal prior distribution for one component of $\boldsymbol{\pi}_{root}$ for different values of k (the distribution of the other components is the same due to symmetry). Since $\boldsymbol{\pi}_{root}|\boldsymbol{\pi} \sim \mathcal{D}(k\boldsymbol{\pi})$, each of the elements of $\boldsymbol{\pi}_{root}$ has a beta distribution

$$\pi_{root,i}|\pi_i \sim \mathrm{Beta}(k\pi_i,\ k(1-\pi_i)),$$

where $\pi_i \sim \text{Beta}(1, 3)$ because $\boldsymbol{\pi} \sim \mathcal{D}(1,\ 1,\ 1,\ 1)$. We use five values of k: $k = 1,\ 4,\ 16,\ 64$, and also the case of a perfect positive dependence between the $\boldsymbol{\pi}_{root}$ and $\boldsymbol{\pi}$ where $k \to \infty$ (denoted by "Stat", because this case corresponds to a stationary model). For smaller values of k the support is given to small values of $\pi_{root,i}$ ($\pi_{root,i} < 0.2$). As k increases, the distribution of $\pi_{root,i}$ approaches the distribution of π_i. The larger the value of k, the more support is given to values of $\pi_{root,i}$ between approximately 0.2 and 0.6, and the less support is given to values of $\pi_{root,i} > 0.6$ (Figure 4.1). We continue our analysis with (i) calculating the marginal prior correlation between $\pi_{root,i}$ and π_i for different values of k, and (ii) performing prior predictive simulations for different values of k.

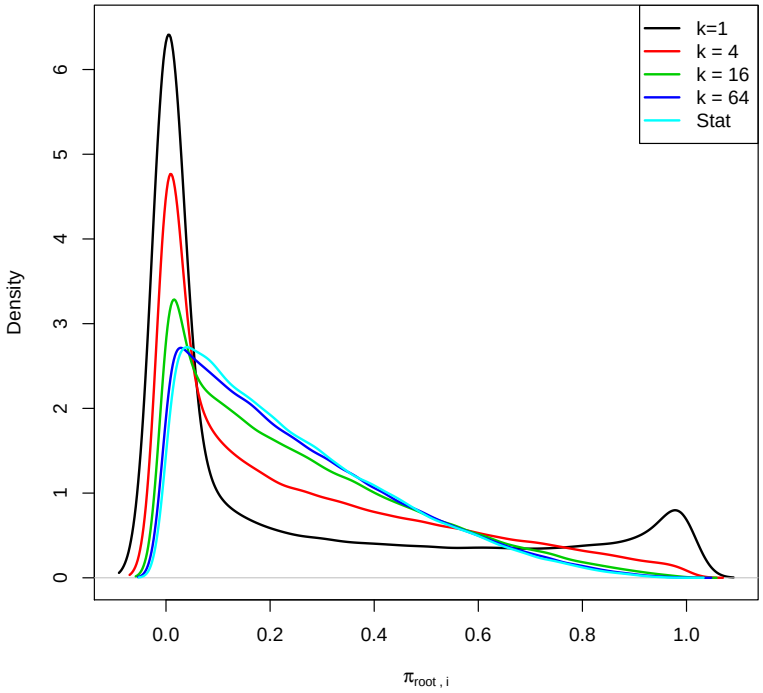


Figure 4.1: Prior distribution for one component of the composition at the root $\boldsymbol{\pi}_{root}$ for different values of the concentration parameter k (the distribution of the other components is the same due to symmetry).

(i) In order to calculate the marginal prior correlation between $\pi_{root,i}$ and π_i, we first calculate the marginal prior variance of $\pi_{root,i}$. Let us denote π_i by π and $\pi_{root,i}$ by π_{root}. The marginal prior variance of π_{root} is

$$\text{Var}(\pi_{root}) = \text{E}\left\{\text{Var}(\pi_{root}|\pi, k)\right\} + \text{Var}\left\{\text{E}(\pi_{root}|\pi, k)\right\}. \tag{4.1}$$

The first term on the RHS of equation (4.1) is

$$\begin{aligned}
\mathrm{E}\left\{\mathrm{Var}(\pi_{root}|\pi,k)\right\} &= \mathrm{E}\left\{\frac{k\pi \times k(1-\pi)}{(k\pi+k(1-\pi))^2(k\pi+k(1-\pi)+1)}\right\} \\
&= \mathrm{E}\left\{\frac{k^2\pi(1-\pi)}{k^2(k+1)}\right\} \\
&= \mathrm{E}\left(\frac{\pi-\pi^2}{k+1}\right) \\
&= \frac{1}{k+1}\left\{\mathrm{E}(\pi)-\mathrm{E}\left(\pi^2\right)\right\} \\
&= \frac{1}{k+1}\left[\mathrm{E}(\pi)-\mathrm{Var}(\pi)-\{\mathrm{E}(\pi)\}^2\right] \\
&= \frac{1}{k+1}(0.25-0.0375-0.0625) \\
&= \frac{0.15}{k+1}.
\end{aligned}$$

The second term on the RHS of equation (4.1) is

$$\begin{aligned}
\mathrm{Var}\left\{\mathrm{E}\left(\pi_{root}|\pi,k\right)\right\} &= \mathrm{Var}\left\{\frac{k\pi}{k\pi+k(1-\pi)}\right\} \\
&= \mathrm{Var}(\pi) = 0.0375.
\end{aligned}$$

Adding both terms on the RHS of equation (4.1) together gives

$$\mathrm{Var}(\pi_{root}) = \frac{0.15}{k+1} + 0.0375.$$

We next calculate the marginal prior covariance between π_{root} and π as

$$\mathrm{Cov}(\pi_{root},\pi) = \mathrm{E}(\pi_{root}\pi) - \mathrm{E}(\pi_{root})\mathrm{E}(\pi). \tag{4.2}$$

The first term on the RHS of equation (4.2) is

$$\mathrm{E}(\pi_{root}\pi) = \mathrm{E}_\pi\{\mathrm{E}(\pi_{root}\pi|\pi)\} = \mathrm{E}_\pi\{\pi\mathrm{E}(\pi_{root}|\pi)\} = \mathrm{E}(\pi^2).$$

Now equation (4.2) can be rewritten as

$$\mathrm{Cov}(\pi_{root},\pi) = \mathrm{E}(\pi^2) - \mathrm{E}(\pi_{root})\mathrm{E}(\pi) = \mathrm{E}(\pi^2) - \{\mathrm{E}(\pi)\}^2 = \mathrm{Var}(\pi).$$

Using the marginal prior covariance $\mathrm{Cov}(\pi_{root}, \pi)$ and marginal prior variance $\mathrm{Var}(\pi_{root})$, we can calculate the the marginal prior correlation between π_{root} and π:

$$\begin{aligned}
\mathrm{Cor}(\pi_{root}, \pi) &= \frac{\mathrm{Cov}(\pi_{root}, \pi)}{\sqrt{\mathrm{Var}(\pi_{root})}\sqrt{\mathrm{Var}(\pi)}} \\
&= \frac{\mathrm{Var}(\pi)}{\sqrt{\mathrm{Var}(\pi_{root})}\sqrt{\mathrm{Var}(\pi)}} \\
&= \sqrt{\frac{\mathrm{Var}(\pi)}{\mathrm{Var}(\pi_{root})}} \\
&= \sqrt{\frac{0.0375}{0.15(k+1)^{-1} + 0.0375}}.
\end{aligned} \tag{4.3}$$

The values of $\mathrm{Cor}(\pi_{root}, \pi)$ for different values of k are

k	$\mathrm{Cor}(\pi_{root}, \pi)$
1	0.5773503
4	0.745356
16	0.887354
64	0.9705818

As $k \to \infty$, $\mathrm{Cor}(\pi_{root}, \pi) \to 1$ and we get perfect positive dependence between $\boldsymbol{\pi}_{root}$ and $\boldsymbol{\pi}$. For $k = 16$ the marginal prior correlation between the π_{root} and π is moderate. This fact combined with the information from the plot shown on Figure 4.1 suggests that $k = 16$ is a sensible choice.

(ii) We perform prior predictive simulations for three different numbers of taxa ($n = 5$, $n = 16$ and $n = 36$). For each number of taxa we use five values of k: $k = 1,\ 4,\ 16$, and also the case of independence between the $\boldsymbol{\pi}_{root}$ and $\boldsymbol{\pi}$ (denoted by "Ind") and the case of a perfect positive dependence between the $\boldsymbol{\pi}_{root}$ and $\boldsymbol{\pi}$ (denoted by "Stat"). Figure 4.2 shows the prior predictive means of the 0-th, 25-th, 50-th, 75-th and 100-th percentiles of one component of the empirical composition in the n-taxa alignment for each value of k and n. For $k = 16$ the empirical composition is modestly concentrated around its mean and is in the range of biologically plausible values, thus confirming our choice of $k = 16$.

Inferring the concentration parameter

In order to infer the concentration parameter we adopt an inverse gamma prior $k \sim \mathrm{IG}(a, b)$. Based on the previous analysis, we chose to centre the prior at the value of 16. In order to choose the values of the hyperparameters a and b, we first analyse the distribution of k for different values of a and b (Figure 4.3a), and also the marginal prior density for one component of the composition at the root $\pi_{root,i}$ for different values of a

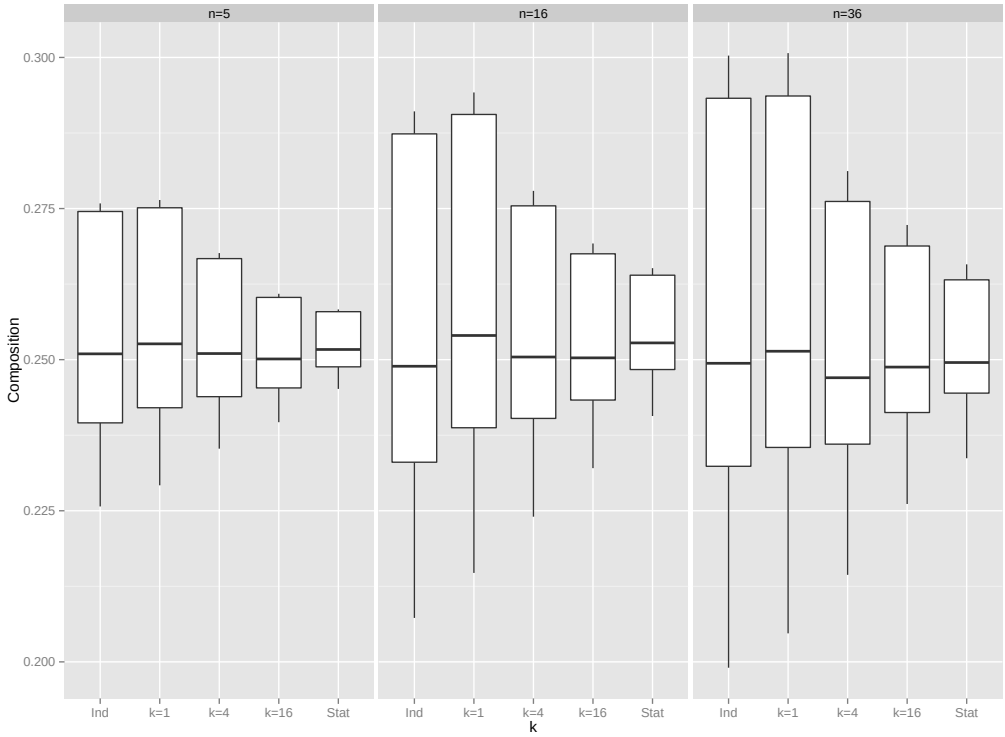
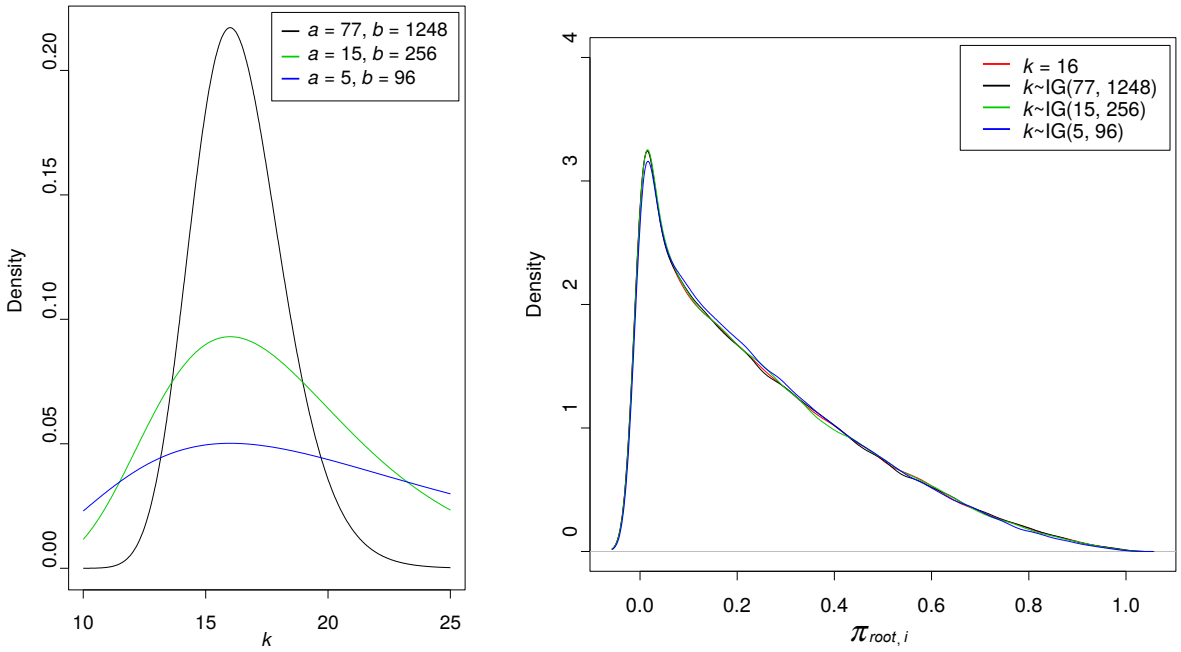


Figure 4.2: Prior predictive means of the 0-th, 25-th, 50-th, 75-th and 100-th percentiles of one component of the empirical composition in the n-taxa alignment for each value of k and n. The cases of independence between the $\boldsymbol{\pi}_{root}$ and $\boldsymbol{\pi}$ is denoted by "Ind" and the case of a perfect positive dependence between the $\boldsymbol{\pi}_{root}$ and $\boldsymbol{\pi}$ is denoted by "Stat".



(a) (b)

Figure 4.3: Graphical analysis of the composition at the root $\boldsymbol{\pi}_{root}$ for different values of the concentration parameter k. (a) Density of k for different values of a and b ($k \sim \text{IG}(a, b)$). In each case k has the mean of 16. (b) Marginal density of one element of the $\boldsymbol{\pi}_{root}$ for different a and b.

Table 4.1: Marginal prior variance of π_{root} for different hyperparameters a and b of the prior for the concentration parameter k.

a	b	$\mathrm{Var}(\pi_{root})$
5	96	0.04485711
15	256	0.04577384
77	1248	0.04621079

Table 4.2: Marginal prior correlation between the composition at the root and the stationary distribution for different hyperparameters a and b of the prior for the concentration parameter k.

a	b	$\mathrm{Cor}(\pi_{root}, \pi)$
5	96	0.9143237
15	256	0.9051217
77	1248	0.9008323

and b (Figure 4.3b). Interestingly, the densities of one element of $\boldsymbol{\pi}_{root}$ are very similar for different a and b, suggesting that the choice of the hyperparameters does not really affect the marginal distribution of the $\boldsymbol{\pi}_{root}$. To confirm this result we calculate the marginal prior variance for one of the elements of $\boldsymbol{\pi}_{root}$ according to equation (4.1)

$$\begin{aligned}
\mathrm{Var}(\pi_{root}) &= \mathrm{E}\left\{Var(\pi_{root}|\pi, k)\right\} + 0.0375 \\
&= 0.15\mathrm{E}\left(\frac{1}{k+1}\right) + 0.0375 \\
&= 0.15\int_0^\infty \frac{b^a}{\Gamma(a)}\frac{\mathrm{e}^{\frac{-b}{k}}}{k+1}k^{-a-1}\mathrm{d}k + 0.0375 \\
&= 0.15\frac{b^a}{\Gamma(a)}\int_0^\infty \frac{\mathrm{e}^{\frac{-b}{k}}}{k+1}k^{-a-1}\mathrm{d}k + 0.0375.
\end{aligned}$$

Table 4.1 shows that the marginal prior variance of one of the elements of the composition at the root is very similar for different hyperparameters a and b; the integrals here were evaluated using http://www.numberempire.com/definiteintegralcalculator.php. The marginal prior correlation between the composition at the root and the stationary distribution, calculated according to equation (4.3) is also very similar for different values of a and b (Table 4.2), in accord with the simulations results.

In conclusion, the analytical results confirm that the composition at the root is not affected by the choice of the hyperparameters of the concentration parameter. We therefore decide to fix the concentration parameter at the value of 16.

4.1.2 Posterior inference via MCMC

The posterior distribution of the unknowns is given by

$$\pi(\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau) \propto \pi(\boldsymbol{\pi}, \kappa, \alpha, \boldsymbol{\ell}, \tau) \times \pi(\boldsymbol{\pi}_{root}|\boldsymbol{\pi}) \times \pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)$$

and the following Metropolis-within-Gibbs algorithm is used to generate posterior samples.
(i) Metropolis-Hastings step for the composition vector $\boldsymbol{\pi}$:

Prior: $\boldsymbol{\pi} \sim \mathcal{D}(\alpha_\pi \boldsymbol{\pi}_0)$, $\alpha_\pi = 4$, $\boldsymbol{\pi}_0 = (0.25,\ 0.25,\ 0.25,\ 0.25)$.
Proposal: $\boldsymbol{\pi}' \sim \mathcal{D}(a_\pi \boldsymbol{\pi})$, where $\boldsymbol{\pi}$ is the current value, a_π is a tuning parameter.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\pi}')\pi(\boldsymbol{\pi}_{root}|\boldsymbol{\pi}')}{\pi(\boldsymbol{\pi})\pi(\boldsymbol{\pi}_{root}|\boldsymbol{\pi})} \times \frac{q(\boldsymbol{\pi}|\boldsymbol{\pi}')}{q(\boldsymbol{\pi}'|\boldsymbol{\pi})} \times \frac{\pi(D|\boldsymbol{\pi}', \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \prod_{i=1}^{4} \frac{\Gamma(a_\pi \pi_i)\Gamma(k\pi_i)}{\Gamma(a_\pi \pi_i')\Gamma(k\pi_i')} \pi_i^{(a_\pi \pi_i' - \alpha_\pi \pi_{0i})} \pi_i'^{(\alpha_\pi \pi_{0i} - a_\pi \pi_i)} \pi_{root,i}^{k(\pi_i' - \pi_i)} \\ &\quad \times \frac{\pi(D|\boldsymbol{\pi}', \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

(ii) Metropolis-Hastings step for the composition at the root $\boldsymbol{\pi}_{root}$:

Prior: $\boldsymbol{\pi}_{root} \sim \mathcal{D}(k\boldsymbol{\pi})$.
Proposal: $\boldsymbol{\pi}'_{root} \sim \mathcal{D}(a_{\pi_r} \boldsymbol{\pi}_{root})$, where $\boldsymbol{\pi}_{root}$ is the current value, a_{π_r} is a tuning parameter.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{\pi}'_{root})}{\pi(\boldsymbol{\pi}_{root})} \times \frac{q(\boldsymbol{\pi}_{root}|\boldsymbol{\pi}'_{root})}{q(\boldsymbol{\pi}'_{root}|\boldsymbol{\pi}_{root})} \times \frac{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}'_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \prod_{i=1}^{4} \frac{\Gamma(a_{\pi_r} \pi_{root,i})}{\Gamma(a_{\pi_r} \pi'_{root,i})} \pi_{root,i}^{(a_{\pi_r} \pi'_{root,i} - k\pi_i)} \pi_{root,i}'^{(k\pi_i - a_{\pi_r} \pi_{root,i})} \\ &\quad \times \frac{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}'_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

(iii) Metropolis-Hastings step for the transition-transversion rate ratio κ:

Prior: $\kappa \sim \mathrm{LN}(\log \kappa_0, \xi^2)$, $\kappa_0 = 1$, $\xi = 0.8$.
Proposal: $\kappa' \sim \mathrm{LN}(\log \kappa, a_\kappa^2)$, where κ is the current value.

Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\kappa')}{\pi(\kappa)} \times \frac{q(\kappa|\kappa')}{q(\kappa'|\kappa)} \times \frac{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa', \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \exp\left[\frac{1}{2\xi^2}\left\{(\log\kappa)^2 - (\log\kappa')^2 + 2\log\kappa_0(\log\kappa' - \log\kappa)\right\}\right] \\ &\quad \times \frac{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa', \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)}. \end{aligned}$$

(iv) Metropolis-Hastings step for the joint move of the root split and the composition at the root $\boldsymbol{\pi}_{root}$:

This step is introduced in order to improve the mixing and the speed of convergence. It is motivated by the fact that changing the root split independently of the composition at the root might result in non-compatibility of the root split and the composition at the root, and therefore in low acceptance rates for the root move. In this move we propose both a new root as described in Section 3.1.4, and a new value for the composition at the root $\boldsymbol{\pi}'_{root} \sim \mathcal{D}(a_{\pi_r}\boldsymbol{\pi}_{root})$, where $\boldsymbol{\pi}_{root}$ is the current value and a_{π_r} is a tuning parameter. The acceptance probability of this move is $\min\{1, A\}$, where

$$A = \frac{\pi(\boldsymbol{\pi}'_{root})\pi(\tau')}{\pi(\boldsymbol{\pi}_{root})\pi(\tau)} \times \frac{\pi(\boldsymbol{\ell}')}{\pi(\boldsymbol{\ell})} \times \frac{q(\boldsymbol{\pi}_{root}|\boldsymbol{\pi}'_{root})}{q(\boldsymbol{\pi}'_{root}|\boldsymbol{\pi}_{root})} \times \frac{w(1-w)}{u(1-u)} \times \frac{\ell_{e_g}}{\ell_{e_a} + \ell_{e_b}} \times \frac{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}'_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau')}{\pi(D|\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \alpha, \boldsymbol{\ell}, \tau)},$$

the variable $u \sim \text{Beta}(2, 2)$ is the auxiliary variable for the proposed root move, $w = \ell_{e_g}/(\ell_{e_a} + \ell_{e_b})$ is the auxiliary variable for the reverse move, ℓ_{e_g} is the length of the proposed rooting edge, ℓ_{e_a} and ℓ_{e_b} are the lengths of the two edges adjacent to the current root (Figure 3.5), and $\ell_{e_g}/(\ell_{e_a} + \ell_{e_b})$ is the Jacobian (Blanquart & Lartillot, 2006).

4.1.3 Simulation study

The simulations aim to investigate the influence of different topologies, different levels of non-stationarity, different alignment lengths and different topological priors on the root inference. We present two blocks of simulations. The first block was performed using data simulated under a random tree with the branch lengths sampled from Ga(2,20) (Figure 4.4). The second block was performed using data simulated under a tree derived from Williams *et al.* (2012) (a rooted version of the unrooted 30-taxon tree used in Chapter 3, Figure 3.9). The placement of the root in this tree corresponds to the eocyte hypothesis (root on branch E_2).

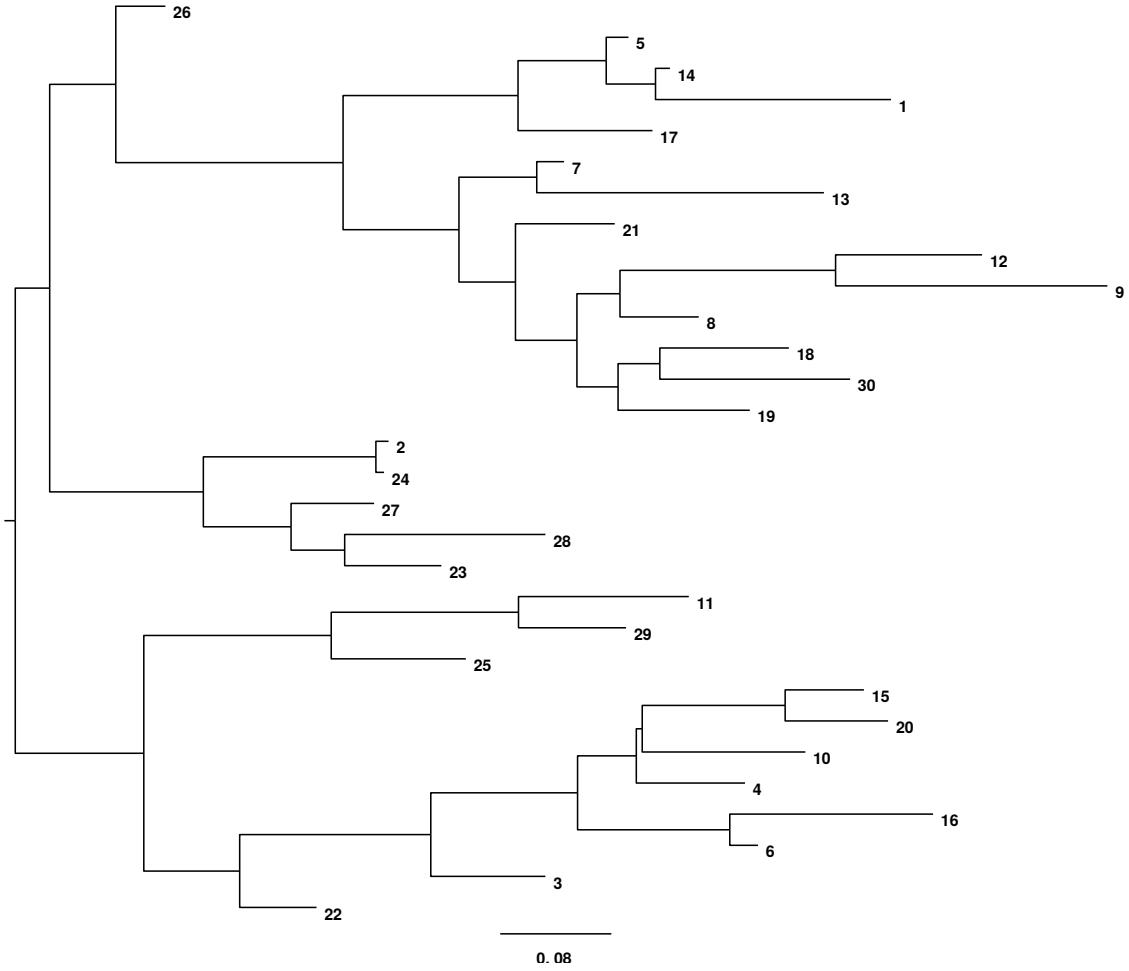


Figure 4.4: Rooted random 30-taxon tree with the branch lengths simulated from Ga(2,20).

Table 4.3: Three data sets simulated with the same HKY85 rate matrix and with different composition at the root $\boldsymbol{\pi}_{root}$.

Data set	$\boldsymbol{\pi}_{root}$
L	(0.27, 0.27, 0.23, 0.23)
M	(0.3, 0.3, 0.2, 0.2)
H	(0.33, 0.33, 0.17, 0.17)

Block One

Different levels of non-stationarity in the data

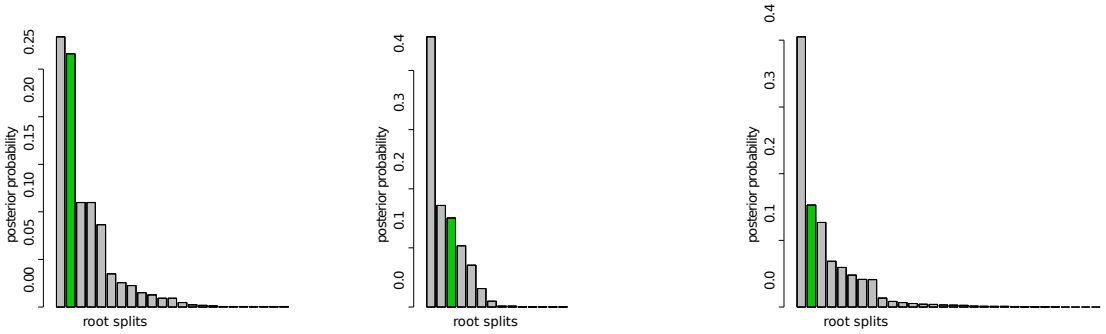
In order to investigate the effect of the non-stationarity on the root inference, we simulated alignments with different levels of non-stationarity under the random tree discussed above. The alignments with the lengths of 2000 sites were simulated with the same HKY85 rate matrix ($\boldsymbol{\pi} = (0.25,\ 0.25,\ 0.25,\ 0.25)$ and $\kappa = 2$), but with different composition vectors at the root vertex (Table 4.3). The data set L has a low level of non-stationarity, the data set M has a moderate level of non-stationarity and the data set H has a high level of non-stationarity. Figure 4.5 shows the posterior distribution of the root splits from three alignments simulated for each level of non-stationarity. The analysis of the alignment with the high level of non-stationarity clearly infers the root better in comparison to the low level of non-stationarity. However, there is a substantial amount of variation between the analyses based on the alignments simulated with the same level of non-stationarity, presumably due to stochastic variation between the simulated data sets.

Block Two

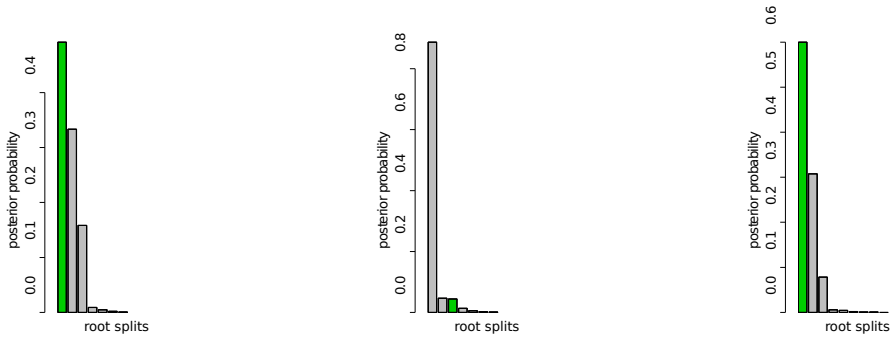
In addition to exploring the effect of different levels of non-stationarity in the data, in this block we explore the effect of different topological priors, different alignment lengths and different concentration parameter for the prior for the composition at the root $\boldsymbol{\pi}_{root}$. Here we use data simulated under a rooted version of the tree shown in Figure 3.9, where the root is placed on branch E_2.

Different topological priors

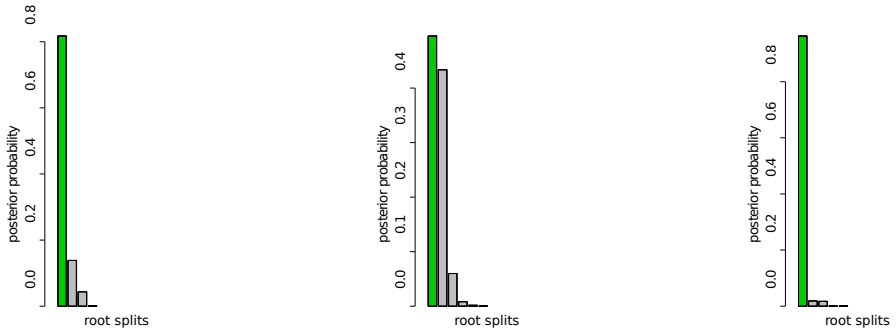
A data set was simulated with a HKY85 rate matrix ($\boldsymbol{\pi} = (0.25,\ 0.25,\ 0.25,\ 0.25)$ and $\kappa = 2$) with the composition at the root $\boldsymbol{\pi}_{root} = (0.25275,\ 0.31256,\ 0.14223,\ 0.29247)$ (simulated from the prior). Figure 4.6 shows the posterior probabilities of the root splits for the analysis with the Yule prior and the structured uniform prior. The posterior probability of the true root is the highest for the Yule prior and the second highest for the structured uniform prior. This suggest that the NS model is able to extract the



(a) L data set

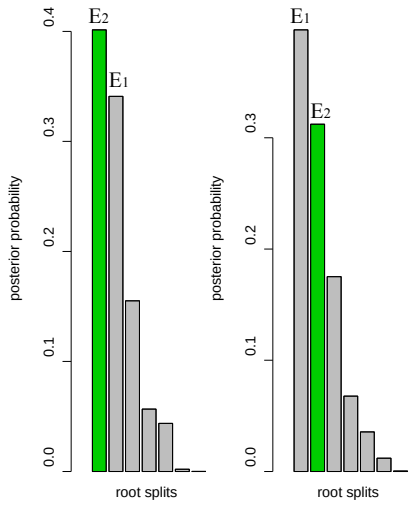


(b) M data set



(c) H data set

Figure 4.5: Posterior distribution of the root splits for the data simulated with different levels of non-stationarity: (a) L data set ($\boldsymbol{\pi}_{root} = (0.27,\ 0.27,\ 0.23,\ 0.23)$); (b) M data set ($\boldsymbol{\pi}_{root} = (0.3,\ 0.3,\ 0.2,\ 0.2)$); (c) H data set ($\boldsymbol{\pi}_{root} = (0.33,\ 0.33,\ 0.17,\ 0.17)$). For each level of non-stationarity three alignments simulated using the same Q matrix are analysed.



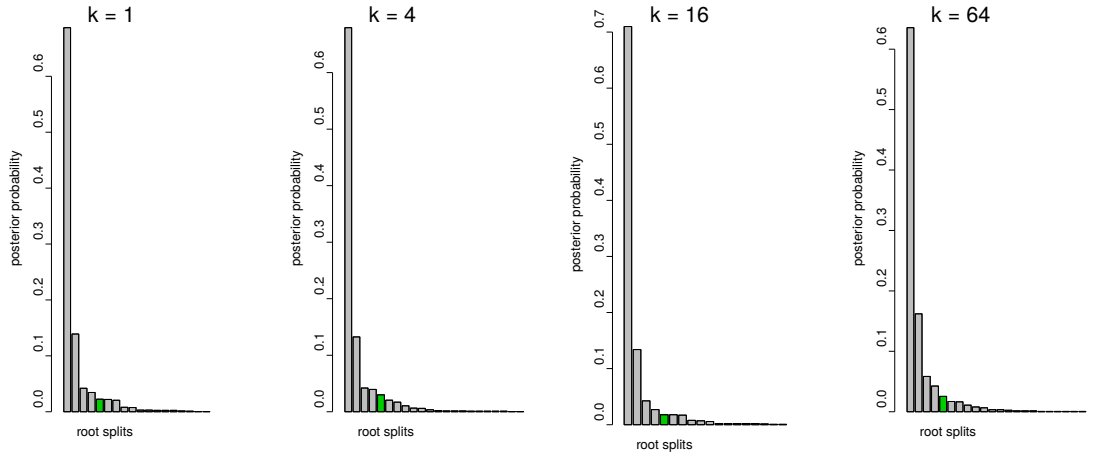
(a) Yule prior (b) Structured uniform prior

Figure 4.6: Posterior distribution of the root splits for the data simulated under the tree shown in Figure 3.9, rooted according to the eocyte hypothesis (root on branch E_2) with $\boldsymbol{\pi}_{root}$ simulated from the prior, analysed with (a) the Yule prior; (b) the structured uniform prior. The true root split has the highest posterior support for the Yule prior and the second highest for the structured uniform prior. On the other hand, the root split on branch E_1 has the highest posterior support for the structured uniform prior and the second highest posterior support for the Yule prior.

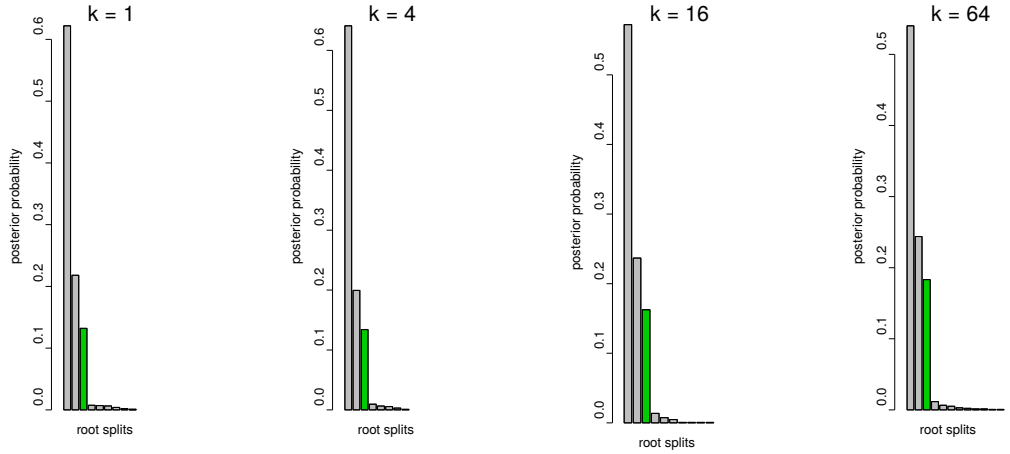
information about the root from data simulated under the tree inferred from the analysis of real data. We note, that this tree is rooted on the rather short branch E_2, thus making the inference more difficult. However, root on the rather long branch E_1 has the highest posterior probability for the structure uniform prior and second highest for the Yule prior. This illustrates the sensitivity to of the model to the topological prior.

Combination of different levels of non-stationarity in the data with different concentration parameters for the prior for the composition at the root

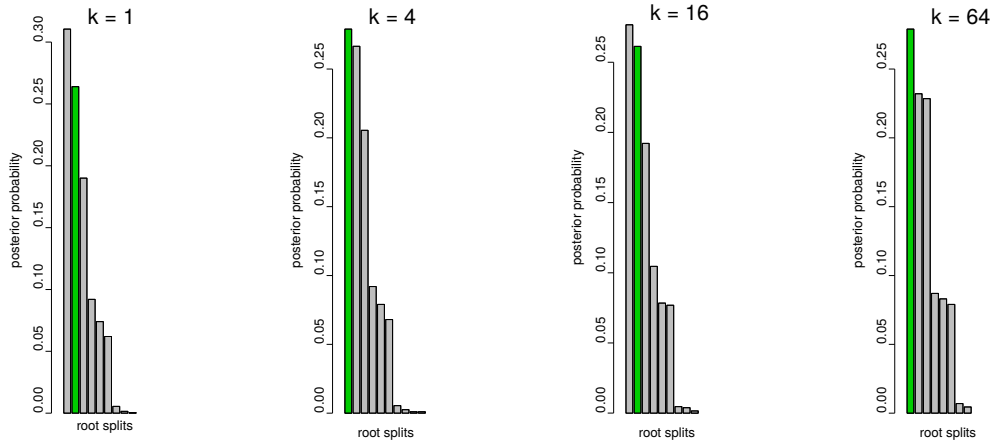
We tested the model on three different data sets having different levels of non-stationarity. The data sets were simulated with the same HKY85 rate matrix ($\boldsymbol{\pi} = (0.25,\ 0.25,\ 0.25,\ 0.25)$ and $\kappa = 2$), but with a different composition at the root vertex (Table 4.3). We already analysed different levels of non-stationarity in the data in block one, however that analysis was performed using only one value of the concentration parameter for the prior for $\boldsymbol{\pi}_{root}$ ($k = 16$). Here we analysed each data set with four values of the concentration parameter for the composition at the root: $k = 1,\ 4,\ 16,\ 64$. Figure 4.7 shows that a large degree of non-stationarity helps to infer the root split better, while the prior for the concentration parameter for the composition at the root makes very little difference. The composition at the root is inferred well in all cases (shown in Figure F.1, Appendix F).



(a) Data set L.



(b) Data set M.

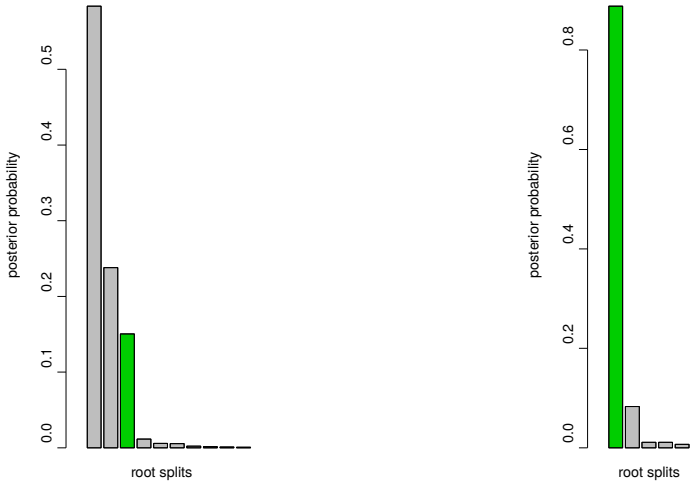


(c) Data set H.

Figure 4.7: Posterior distribution of the root splits for the data simulated under the tree shown in Figure 3.9, with different levels of non-stationarity, analysed with different values of the concentration parameter for the composition at the root.

Different alignment length

We compared the root inference for two data sets simulated with the same parameters but with different lengths. Both data sets were simulated with the moderate level of non-stationarity ($\boldsymbol{\pi}_{root} = (0.3,\ 0.3,\ 0.2,\ 0.2)$), one having 2000 sites, the other having 10000 sites. Increasing the alignment lengths substantially improved the root inference (Figure 4.8). Figure 4.9 shows the posterior distribution of the composition at the root for both alignments. The true values of the composition at the root are shown with the black line. Clearly the composition at the root is better inferred for the longer alignment.



(a) 2000 sites (b) 10000 sites

Figure 4.8: Posterior distribution of the root splits for the alignments with different lengths: (a) 2000 sites; (b) 10000 sites.

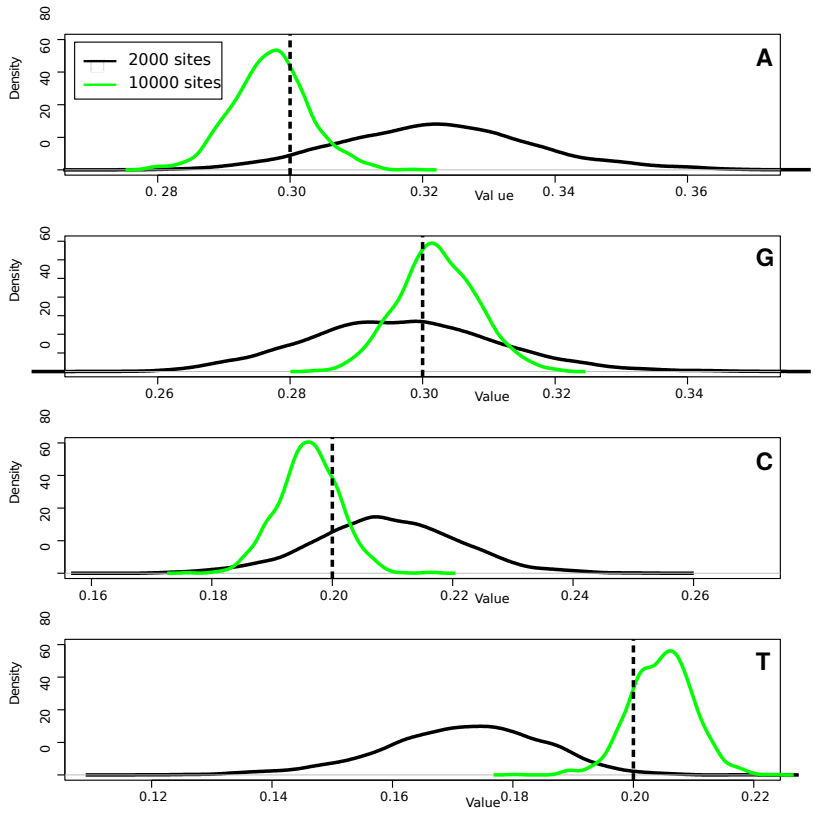


Figure 4.9: Posterior distribution of the composition at the root for the alignments with different lengths (black line - 2000 sites, green line - 10000 sites). The true values are indicated with dashed vertical lines (the same for both alignments).

4.2 Non-stationary non-reversible one component model

The NS model discussed above utilises a reversible HKY85 rate matrix to describe the substitution process along branches. In this section we combine the idea of non-stationarity with the NR model to obtain a non-reversible and non-stationary model which we denote NRNS. We note that in the NRNS model the prior for the composition at the root $\boldsymbol{\pi}_{root}$ depends on the theoretical stationary distribution $\boldsymbol{\pi}_Q$ which is not the same as $\boldsymbol{\pi}$, the composition vector in the underlying HKY85 rate matrix Q^H.

4.2.1 Posterior inference via MCMC

The posterior distribution of the unknowns is given by

$$\pi(\boldsymbol{\pi}, \boldsymbol{\pi}_{root}, \kappa, \sigma, Q, \alpha, \boldsymbol{\ell}, \tau) \propto \pi(Q|\boldsymbol{\pi}, \kappa, \sigma) \times \pi(\boldsymbol{\pi}, \kappa, \sigma, \alpha, \boldsymbol{\ell}, \tau) \times \pi(\boldsymbol{\pi}_{root}|\boldsymbol{\pi}_Q) \times \pi(D|Q, \boldsymbol{\pi}_{root}, \alpha, \boldsymbol{\ell}, \tau).$$

The Metropolis-within-Gibbs algorithm used to generate posterior samples from the underlying HKY85 rate matrix Q^H is similar to those described previously.
(i) Metropolis-Hastings step for the composition vector $\boldsymbol{\pi}$:

Prior: $\boldsymbol{\pi} \sim \mathcal{D}(\alpha_\pi \boldsymbol{\pi}_0)$, $\alpha_\pi = 4$, $\boldsymbol{\pi}_0 = (0.25,\ 0.25,\ 0.25,\ 0.25)$.
Proposal: $\boldsymbol{\pi}' \sim \mathcal{D}(a_\pi \boldsymbol{\pi})$, where $\boldsymbol{\pi}$ is the current value, a_π is a tuning parameter.
Acceptance probability: $\min\{1, A\}$, where

$$A = \frac{\pi(\boldsymbol{\pi}')}{\pi(\boldsymbol{\pi})} \times \frac{q(\boldsymbol{\pi}|\boldsymbol{\pi}')}{q(\boldsymbol{\pi}'|\boldsymbol{\pi})} \times \frac{\pi(Q|\boldsymbol{\pi}', \kappa, \sigma)}{\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)} \times \frac{\pi(D|Q, \boldsymbol{\pi}_{root}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \boldsymbol{\pi}_{root}, \alpha, \boldsymbol{\ell}, \tau)}.$$

Note that A is calculated similarly to its analogue in the NR model.

In fact, Metropolis-Hastings steps for κ, σ and α are also similar to their analogues in the NR model, and the Metropolis-Hastings step for the composition at the root is similar to its analogue in the NS model. However, the Metropolis-Hastings step for the off-diagonal elements of the rate matrix Q is different from its analogue in the NR model. This step includes evaluating the prior for the distribution at the root, since the latter depends on the rate matrix Q:

Prior: $q_{ij} \sim \mathrm{LN}(\log q_{ij}^H, \sigma^2)$.
Proposal: $q'_{ij} \sim \mathrm{LN}(\log q_{ij},\ a_q^2)$, where q_{ij} is the current value.
Acceptance probability: $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(Q'_{ij}|\boldsymbol{\pi}, \kappa, \sigma)\pi(\boldsymbol{\pi}_{root}|\boldsymbol{\pi}_{Q'})}{\pi(Q_{ij}|\boldsymbol{\pi}, \kappa, \sigma)\pi(\boldsymbol{\pi}_{root}|\boldsymbol{\pi}_{Q})} \times \frac{q(q_{ij}|q'_{ij})}{q(q'_{ij}|q_{ij})} \times \frac{\pi(D|Q', \boldsymbol{\pi}_{root}, \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \boldsymbol{\pi}_{root}, \alpha, \boldsymbol{\ell}, \tau)} \\ &= \prod_{i=1}^{4} \frac{\Gamma(k\pi_{Q,i})}{\Gamma(k\pi_{Q',i})} \pi_{root,i}^{k(\pi_{Q',i} - \pi_{Q,i})} \\ &\quad \times \exp\left[\frac{1}{2\sigma^2} \sum_{i \neq j} \left\{(\log q_{ij})^2 - (\log q'_{ij})^2 + 2 \log q_{ij}^H (\log q'_{ij} - \log q_{ij})\right\}\right] \\ &\quad \times \frac{\pi(D|Q', \boldsymbol{\pi}_{root}(Q'), \alpha, \boldsymbol{\ell}, \tau)}{\pi(D|Q, \boldsymbol{\pi}_{root}(Q), \alpha, \boldsymbol{\ell}, \tau)}, \end{aligned}$$

and $\boldsymbol{\pi}_{Q'}$ is the stationary distribution obtained from Q'.

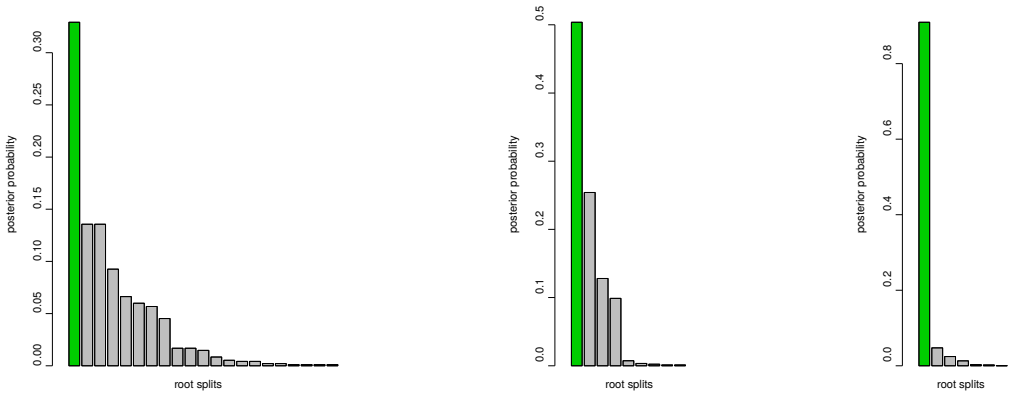
4.2.2 Simulation study

The simulations focus on investigating the effect of different levels of non-stationarity and non-reversibility in the data simulated under the trees used in the simulations for the NS model (Section 4.1.3).

Block One

Different levels of non-stationarity

We use a random 30-taxon tree with branch lengths simulated from Ga(2,20) (Figure 4.4) to simulate alignments with different levels of non-stationarity. The alignments contain the same (moderate) level of non-reversibility, each alignment has 2000 sites. Figure 4.10 shows the posterior distribution of the root splits for 3 alignments simulated with $\sigma = 0.1$ and different levels of non-stationarity (Table 4.3). As expected, the inference is better for the data set having a higher level of non-stationarity. Apart from higher posterior support for the true root split, the analysis of the alignment with larger degree of non-stationarity also shows less posterior variation on the root splits.



(a) data set L (b) data set M (c) data set H

Figure 4.10: Posterior distribution for the root splits for the data simulated with $\sigma = 0.1$ and different levels of non-stationarity (a) L data set ($\boldsymbol{\pi}_{root} = (0.27,\ 0.27,\ 0.23,\ 0.23)$); (b) M data set ($\boldsymbol{\pi}_{root} = (0.3,\ 0.3,\ 0.2,\ 0.2)$); (c) H data set ($\boldsymbol{\pi}_{root} = (0.33,\ 0.33,\ 0.17,\ 0.17)$).

Different levels of non-stationarity and non-reversibility

This analysis comprises 15 data sets simulated with different levels of non-reversibility (low: $\sigma = 0.05$, moderate: $\sigma = 0.1$, high: $\sigma = 0.3$) and different degrees of non-stationarity (low: $\boldsymbol{\pi}_{root} = (0.27, 0.27, 0.23, 0.23)$, moderate: $\boldsymbol{\pi}_{root} = (0.3, 0.3, 0.2, 0.2)$, high: $\boldsymbol{\pi}_{root} = (0.33, 0.33, 0.17, 0.17)$) (Table 4.4). The stationary case is denoted by "Stat" ($\boldsymbol{\pi}_{root} = (0.25, 0.25, 0.25, 0.25)$), and the reversible case is denoted by "Rev" ($\sigma = 0$). Notice, that the case of Stat and Rev corresponds to a HKY85 model (not analysed because the likelihood does not depend on the root position). The posterior probabilities of the root splits for the 15 data sets are shown on Figure 4.11.

Increasing the level of non-reversibility for alignments simulated under stationary models, and increasing the level of non-stationarity for alignments simulated under reversible models improves the root inference. However, the combination of different levels of non-

Table 4.4: data set with different values of perturbation component and different degrees of non-reversibility.

	Stat	NS (low)	NS (moderate)	NS (high)
Rev	HKY85	NS(l)	NS(m)	NS(h)
NR (low)	NR(l)	NR(l)NS(l)	NR(l)NS(m)	NR(l)NS(h)
NR (moderate)	NR(m)	NR(m)NS(l)	NR(m)NS(m)	NR(m)NS(h)
NR (high)	NR(h)	NR(h)NS(l)	NR(h)NS(m)	NR(h)NS(h)

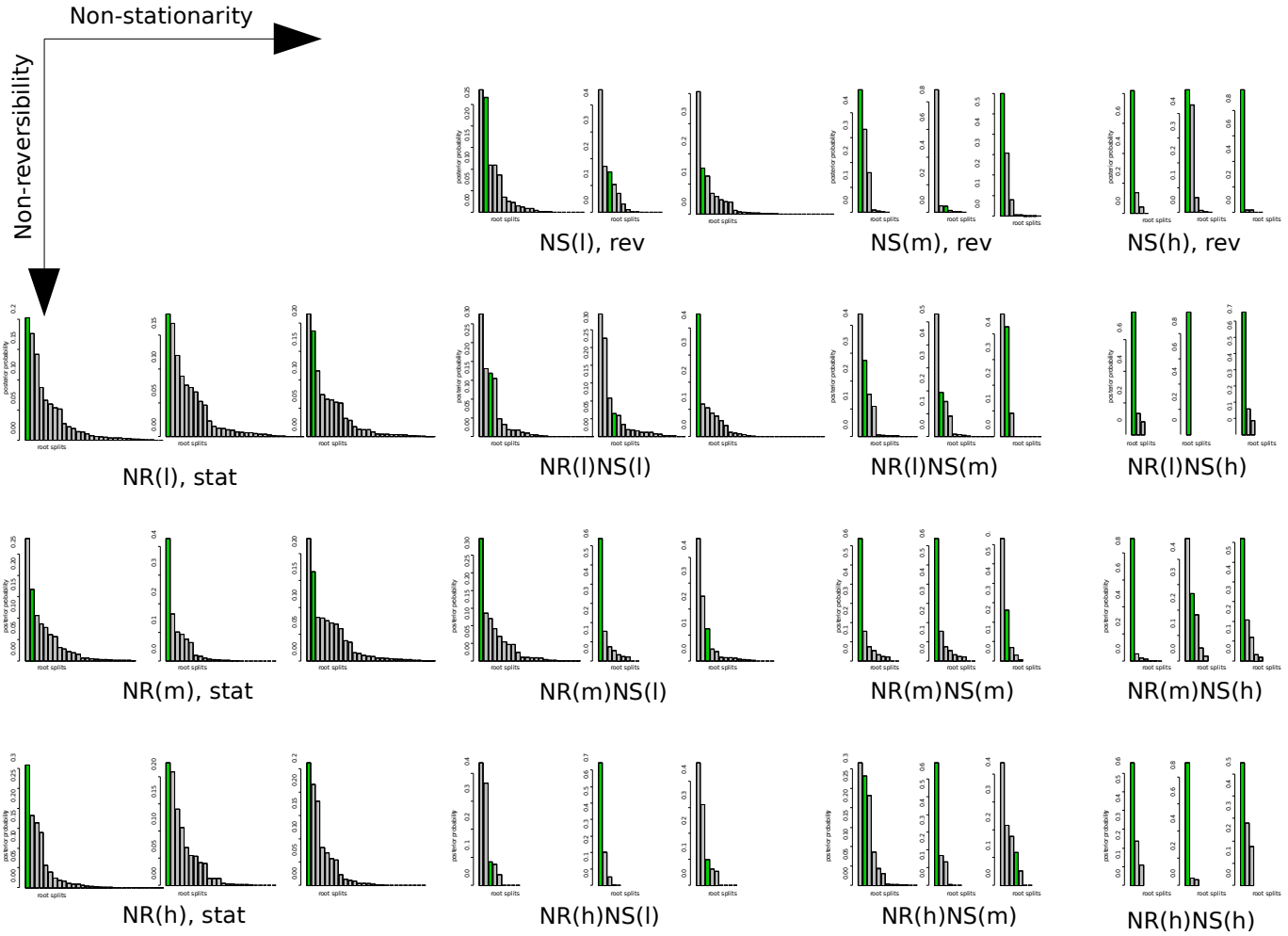


Figure 4.11: Posterior probabilities of the root splits for different degrees of non-reversibility and non-stationarity (Table 4.4). For each case three independent alignments are analysed.

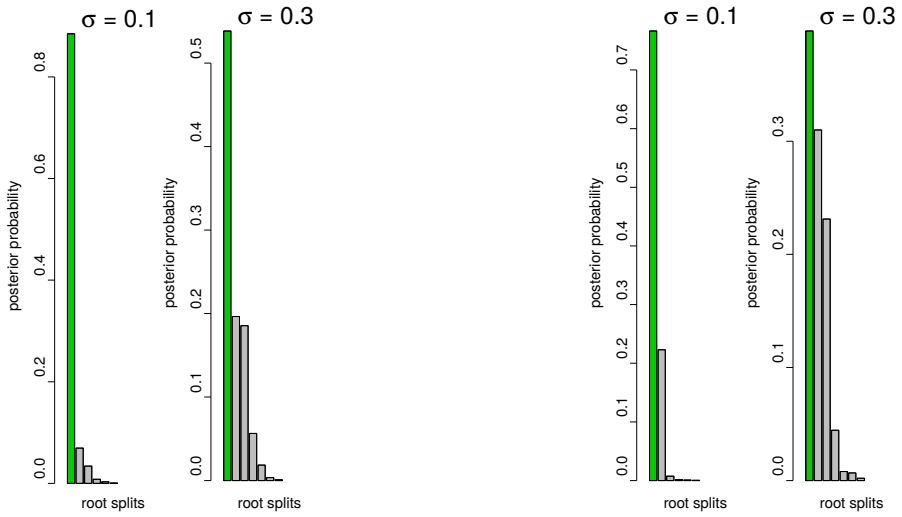
reversibility with non-stationarity does not necessarily do this, probably because the effects of non-reversibility and non-stationarity are confounded. Amongst the non-reversible and non-stationary data sets the root inference is better for the moderate amount of non-reversibility ($\sigma = 0.1$). Increasing the amount of non-stationarity increases the posterior probability of the true root for all degrees of non-reversibility. This suggests that the

signal of non-stationarity is stronger than the signal of non-reversibility.

Block Two

In this block we use the tree shown in Figure 3.9, rooted on edge E_2. Two data sets were simulated with the same HKY85 rate matrix ($\boldsymbol{\pi} = (0.25,\ 0.25,\ 0.25,\ 0.25)$ and $\kappa = 2$) and the same composition at the root $\boldsymbol{\pi}_{root} = (0.25275,\ 0.31256,\ 0.14223,\ 0.29247)$ (simulated from the prior), but different perturbation parameters ($\sigma = 0.1$ and $\sigma = 0.3$). Figure 4.12 shows the posterior distribution for the root splits analysed with two topological priors (the Yule prior and the structured uniform prior). The true root split is inferred as the mode for both values of σ confirming that the effect of non-stationarity seems to dominate over the effect of non-reversibility. Overall this analysis suggests that the NRNS model can extract information about the root for non-reversible and non-stationarity data simulated under the tree inferred from the analysis of real data.

Similarly to the NRNS model, the idea of non-stationarity can be combined with the NR2 model from Chapter 3, thus giving a two component non-reversible and non-stationary model. However, for brevity it is not considered further.



(a) Yule prior (b) Structured uniform prior

Figure 4.12: Posterior distribution of the root splits for the data simulated under the tree shown in Figure 3.9 (rooted on edge E_2) with the NRNS model and with different values of the perturbation parameter ($\sigma = 0.1$ and $\sigma = 0.3$), analysed with (a) the Yule prior; (b) the structured uniform prior.

Summary

In this chapter we have presented two non-stationary models in which the composition at the root is centered on the stationary composition. While the first model (the NS model)

uses a reversible HKY85 rate matrix to describe the substitution of nucleotides, the second model (the NRNS model) utilises a rate matrix of the non-reversible NR model presented in Chapter 3. For each model we performed a simulation study for data containing different degrees of non-stationarity, and in the case of the NRNS model also different degrees of non-reversibility. We found that increasing the level of non-stationarity leads to better root inference for both models, and so does increasing the level of non-reversibility for the NRNS model. However, an analysis of the data with combination of different levels of non-stationarity and non-reversibility showed the potential problem of confounding between the two signals. Analysing data with the same level of non-stationarity and different levels of non-reversibility showed that the effect of non-stationarity dominates the effect of non-reversibility. We also investigated the influence of the length of the sequence alignment to the root inference and found that both the root split and the composition at the root are inferred better for longer alignments.

Chapter 5

Application to experimental data

In this chapter we analyse real biological data sets with non-reversible and non-stationary models: the palaeopolyploid yeasts data, the primates data and the tree of life data. For the yeasts and the primates data there is a robust biological opinion about the position of the root, whereas the root of the tree of life is still an open question in biology. We address the difference in the results obtained with different models and identify possible biological reasons for it, e.g. variation in composition of nucleotides. We show that while non-reversible models are able to extract some information about the root, modelling non-stationarity with just two composition vectors can be misleading for certain data sets.

5.1 Rooting the radiation of palaeopolyploid yeasts

MCMC implementation

In this chapter, all results are based on (almost) un-autocorrelated posterior samples of size at least 5K. These samples were obtained by running the algorithm for at least 1000K iterations, discarding at least 300K iterations as burn-in and then thinning by taking every 100-th iterate to remove autocorrelation. Convergence was diagnosed using the procedure described in Section 2.6.3. This involved initialising two MCMC chains at different starting points and graphically comparing the chains through properties based on model parameters and the relative frequencies of sampled clades. In all cases, the graphical diagnostics gave no evidence of any lack of convergence.

5.1.1 Non-reversible models

We investigated the performance of the non-reversible and non-stationary models on a real biological data set for which there is broad biological consensus on the root position (Byrne & Wolfe, 2005; Hedtke *et al.*, 2006). The lineage leading to *Saccharomyces cerevisiae* (brewer's yeast) and its relatives underwent a conserved whole-genome duplication

(WGD) about 100 million years ago (Wolfe & Shields, 1997; Kellis *et al.*, 2004). Evidence for this WGD, in the form of duplicated genes and genomic regions, is shared by all post-WGD yeasts and defines the group as a clade from which the root of the *Saccharomycetales* is excluded. Current biological opinion on the rooted phylogeny of these species is summarised in Figure 5.1 (Byrne & Wolfe, 2005; http://ygob.ucd.ie, 2015). The

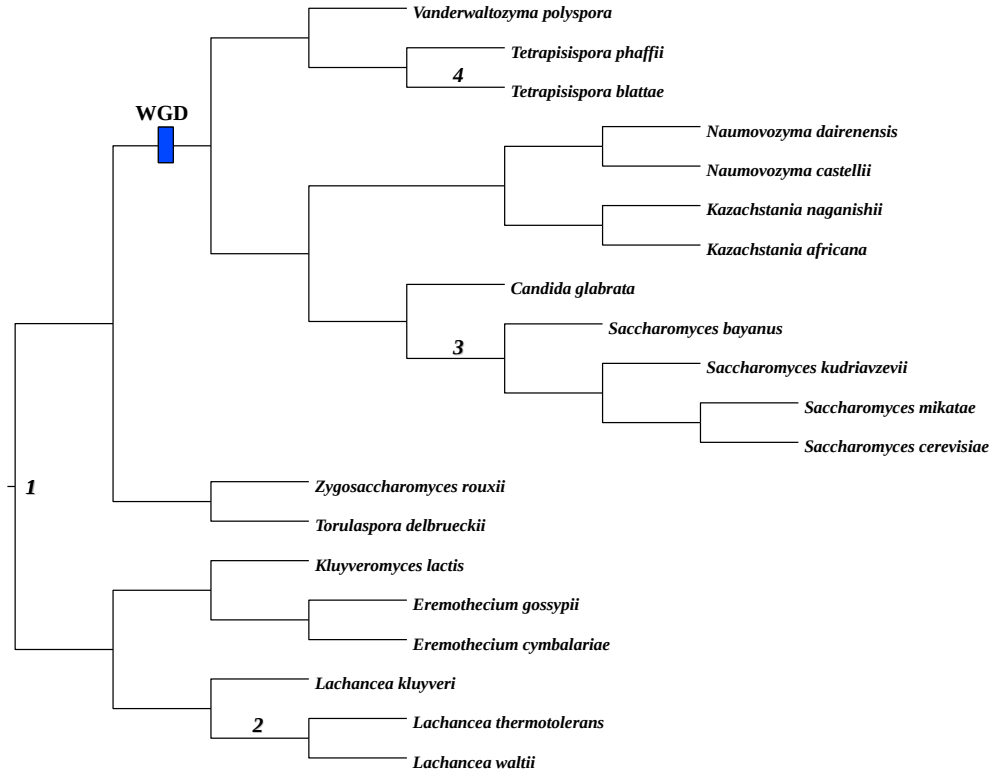
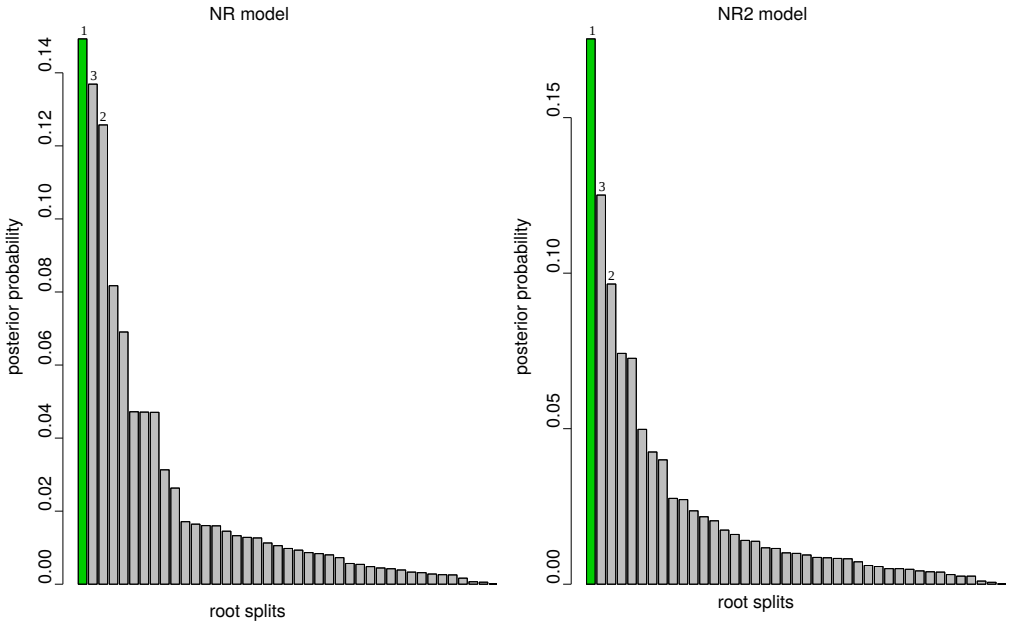


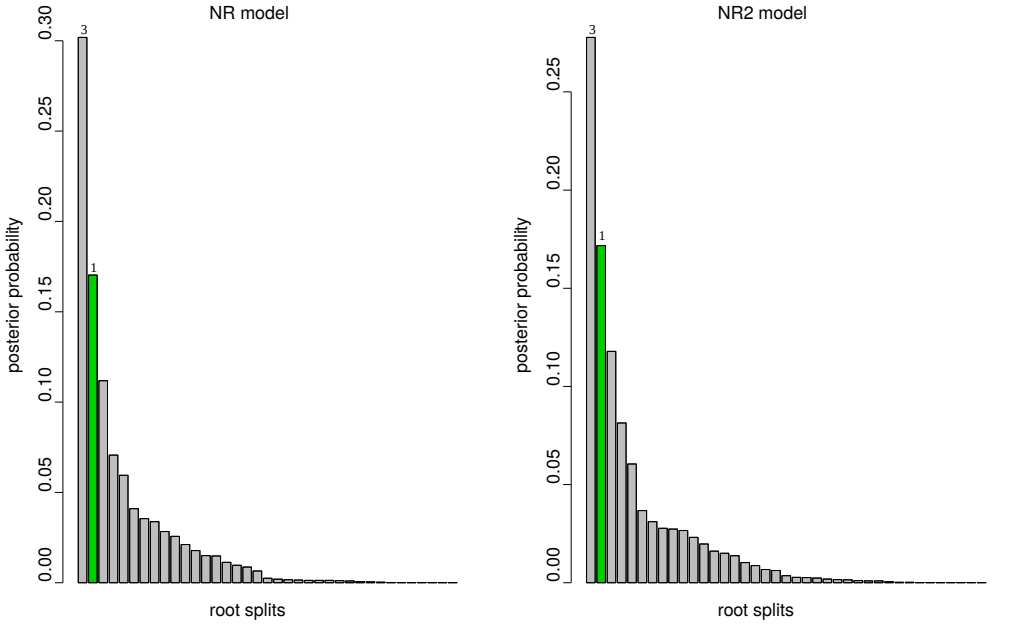
Figure 5.1: Rooted phylogeny of the palaeopolyploid yeasts supported by the whole-gene duplication analysis (not drawn to scale), reproduced from the YGOB website (Byrne & Wolfe, 2005; http://ygob.ucd.ie, 2015). Four different roots indicated by numbers 1 - 4 were inferred in the analysis with the non-reversible and non-stationary models. Root 1 which represents the biologically plausible root was inferred after fitting the GTR model via maximum likelihood (Hedtke *et al.*, 2006).

analysis which generated this tree was based on outgroup rooting (see Section 1.3.1) after fitting the GTR model by maximum likelihood (Hedtke *et al.*, 2006). The root on this tree separates a clade comprising *Eremothecium gossypii, Eremothecium cymbalariae, Kluyveromyces lactis, Lachancea kluyveri, Lachancea thermotolerans* and *Lachancea waltii* from the other species. It is consistent with the timing of the WGD event.

We analysed an alignment of concatenated large and small subunit ribosomal DNA sequences for 20 species of yeast, with a combined length of 4460 sites. The sequences were aligned with MUSCLE (Edgar, 2004), and poorly-aligned regions were detected and removed using TrimAl (Capella-Gutiérrez *et al.*, 2009). Figure 5.2 (a) shows the posterior



(a) Structured uniform prior



(b) Yule prior

Figure 5.2: The posterior distribution of the root splits of the palaeopolyploid yeasts data set for both NR and NR2 models analysed with (a) the structured uniform prior and (b) the Yule prior. Different bars on the plot represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). The roots are mapped in Figure 5.1. In (a) the root split supported by outgroup rooting (Hedtke at al. 2006) has the highest posterior probability (root 1, highlighted). Root 2 is placed within the outgroup and root 3 is placed within the post-WGD clade. In (b) the root split supported by outgroup rooting (Hedtke at al. 2006) has the second highest posterior probability (root 1, highlighted).

distribution of the root splits for both the NR and NR2 models implemented with the structured uniform prior. The root split supported by outgroup rooting (Hedtke *et al.*, 2006) has the highest posterior probability (root 1 in Figure 5.1) for both models. However, there is a substantial amount of uncertainty represented by the non-negligible posterior probabilities of the other root splits. While the third most plausible root is placed within the outgroup (root 2 in Figure 5.1), the second most plausible root is located within the post-WGD clade (root 3 in Figure 5.1). This degree of uncertainty is also reflected in the sensitivity of the analysis to the topological prior: while the structured uniform prior recovered the root supported by the outgroup analysis with the highest posterior support, the Yule prior instead recovered this root with the second-highest support (root 1 in Figure 5.2 (b)). The most plausible root inferred with the Yule prior is placed within the post-WGD clade (root 3 in Figure 5.1), which contradicts the WGD analysis.

The posterior for the non-reversibility parameter σ in the NR model is suggestive of a substantial degree of non-reversibility in the data.This is illustrated in Figure 5.3 which shows posterior density for σ for five data sets analysed in this chapter (the primates and the tree of life data sets will be introduced later in this chapter). For the yeasts data set, it offers no support for values of σ around zero. For some simulated data, we were able to infer the true root with higher posterior support and less uncertainty for such σ. This suggests the presence of other features of the data not accounted for by the model that are masking the root signal.

The rooted majority rule consensus trees from the analyses with the two topological priors are depicted in Figure 5.4. The unrooted topologies of both consensus trees differ

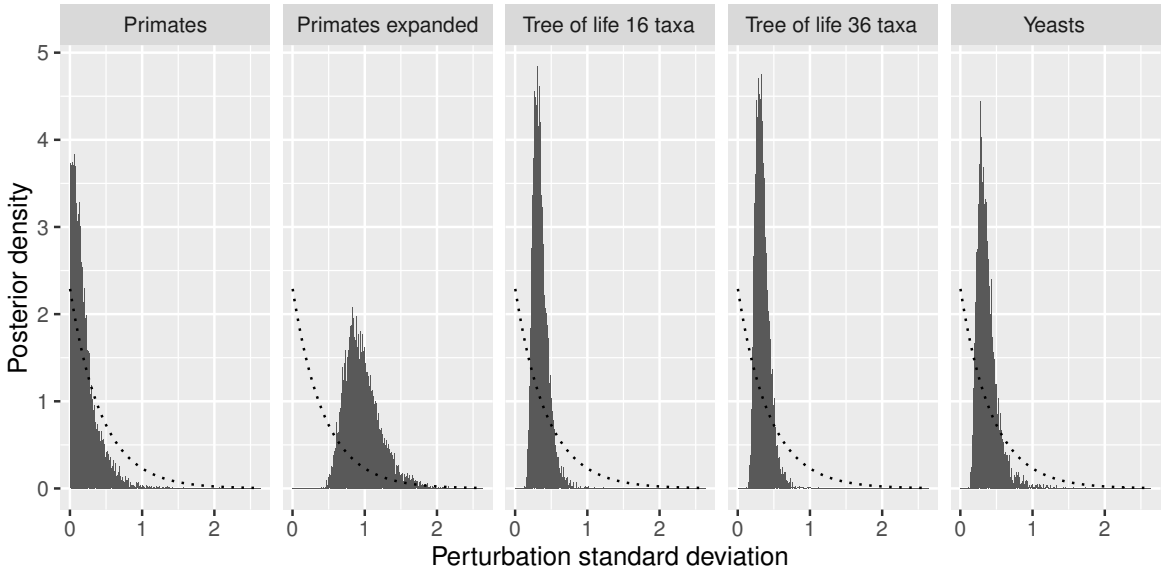
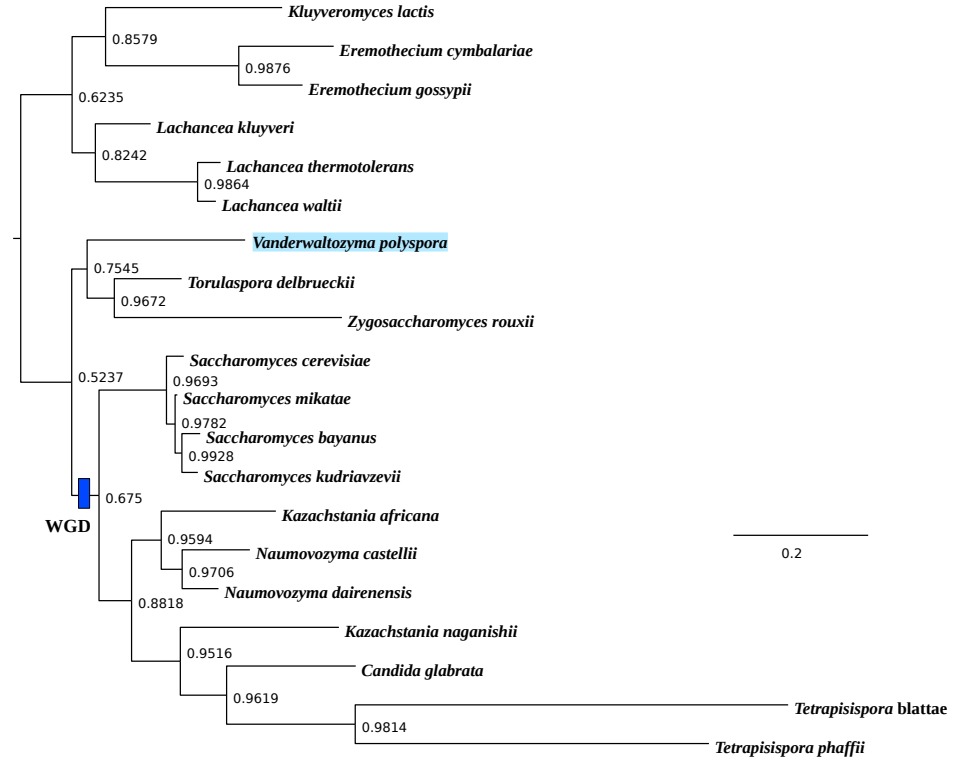
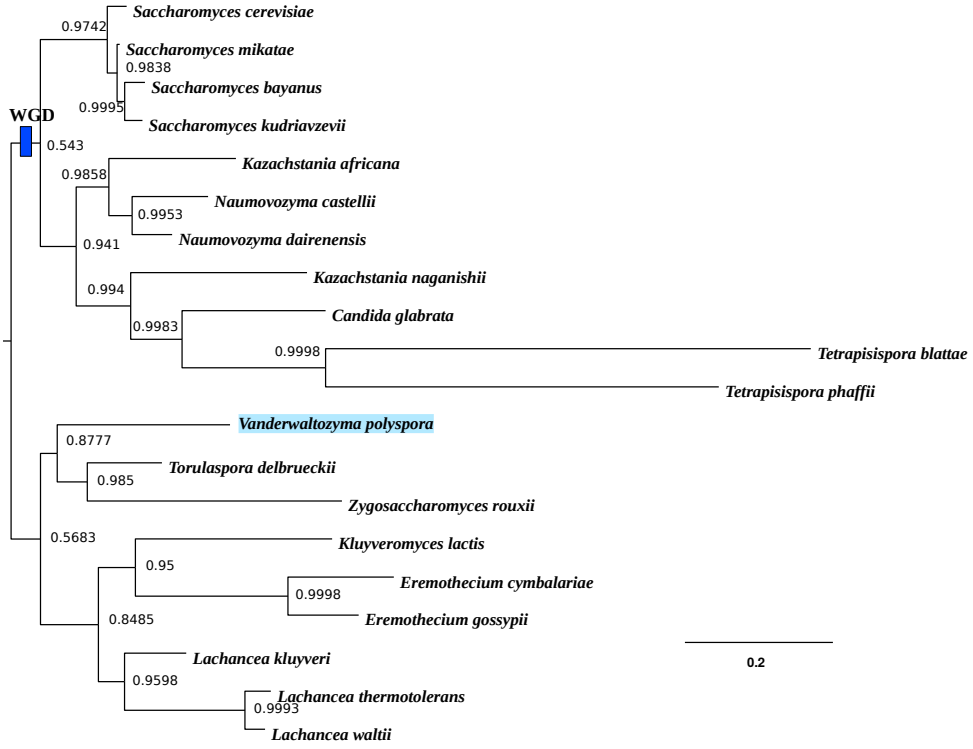


Figure 5.3: Posterior density for the perturbation standard deviation σ for five data sets analysed in this chapter. In each plot, the dotted line represents the prior density for σ.



(a)



(b)

Figure 5.4: Rooted majority rule consensus tree of the palaeopolyploid yeasts data set, inferred under the NR model using (a) the structured uniform prior and (b) the Yule prior, with the WGD event mapped. The trees differ from that supported by the WGD analysis by the placement of *Vanderwaltozyma polyspora* (shaded in blue) within the pre-WGD clade.

from that supported by the WGD by the placement of *Vanderwaltozyma polyspora*. While the WGD analysis places it within the post-WGD clade, in our analysis this taxon is located within the pre-WGD clade. Interestingly, this result is consistent with an analysis performed with the CAT-GTR model (Lartillot & Philippe, 2004), which is a Dirichlet process mixture model accounting for heterogeneity in composition across sites (Section 2.5). It has been shown that the CAT-GTR model can provide a better fit to the data than the site homogeneous models (Cox *et al.*, 2008). On the tree inferred with the CAT-GTR model, *Vanderwaltozyma polyspora* is excluded from the post-WGD clade (Figure 5.5). The placement of *Vanderwaltozyma polyspora* outside the WGD clade is surprising given that the genome of *Vanderwaltozyma polyspora* preserves evidence of having undergone WGD (Scannell *et al.*, 2007).

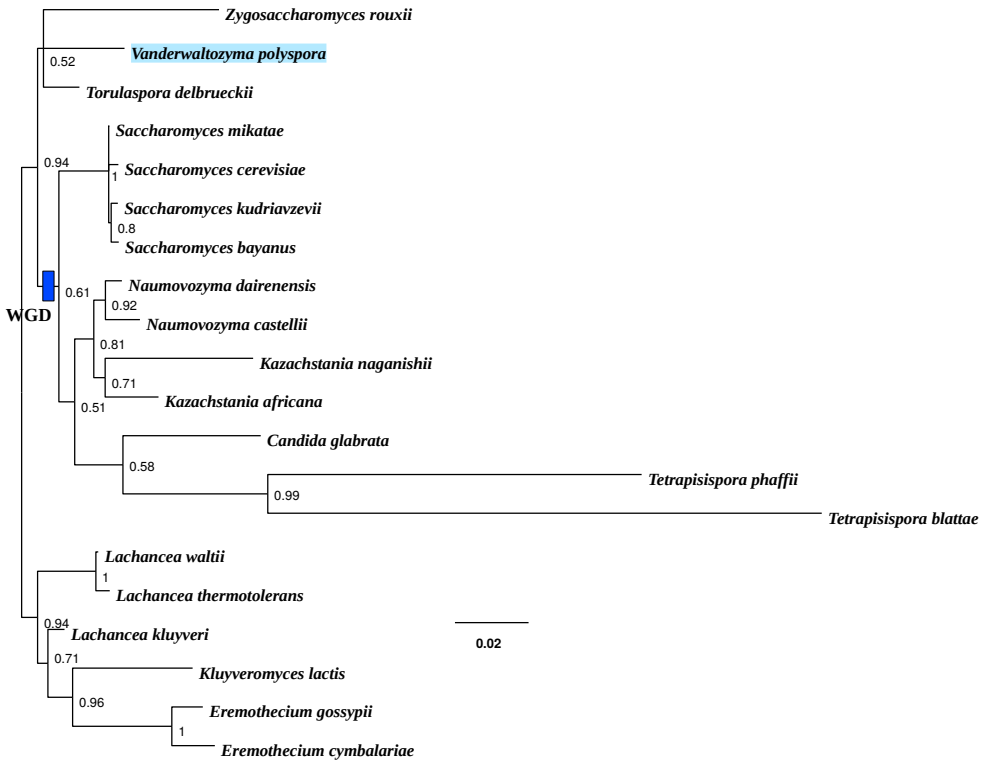


Figure 5.5: Unrooted consensus tree of the palaeopolyploid yeasts data set, inferred with the CAT-GTR model (Lartillot & Philippe, 2004), with the WGD event mapped. Similarly to the NR model, the CAT-GTR model places the *Vanderwaltozyma polyspora* (shaded in blue) within the pre-WGD clade which contradicts the WGD analysis.

This result requires further investigation. However, the similarity between the unrooted consensus trees obtained with the CAT-GTR model and with our non-reversible models suggests that the non-reversible models can not only extract meaningful information about the root position, but also capture information for inferring the unrooted

topology.

It is worth noting that the root split on the majority rule consensus tree (Figure 5.4b) does not match the marginal posterior modal root split (Figure 5.2b). This can happen because the consensus tree is a conditional summary, computed recursively from the leaves to the root, which depends upon the plausibility of sub-clades. On the other hand, the posterior over root splits is a marginal summary which averages over the relationships expressed elsewhere in the tree; see Appendix G for an illustrative example.

5.1.2 Non-stationary models

We also analysed the yeasts data set with the NS and NRNS models to investigate if modelling non-stationarity improved the root inference. Surprisingly, both models (analysed with both topological priors) recovered the root on a pendant edge leading to *Tetrapisispora blattae* with posterior probability of 1 (root 4 in Figure 5.1). This root is located within the post-WGD clade and hence it contradicts current biological opinion. In order to investigate this result we analysed the empirical composition of nucleotides. Since the composition vector $\boldsymbol{\pi}$ is defined on the four-dimensional simplex, its graphical representation is not straightforward. To provide a graphical visualisation of the composition we therefore transformed each composition vector $\boldsymbol{\pi}$ to the three-dimensional real parameter $\boldsymbol{\beta}$. The transformation was achieved by applying a multinomial logit reparametrisation followed by a linear mapping, to obtain three unconstrained real numbers corresponding to each composition vector, as described in Heaps *et al.* (2014). The procedure consists of two parts:
(i) Multinomial logit reparametrisation of the (empirical) composition vector $\boldsymbol{\pi}_j$ for species j:

$$\pi_{jk} = \frac{\exp(\alpha_{jk})}{\sum_{m=1}^{4} \exp(\alpha_{jm})},$$

where $j = 1, \ldots, n$ are the species, and $k = 1, \ldots, 4$ are the nucleotides. Here $\alpha_{jk} \in \mathbb{R}^4$ and $\sum_{k=1}^{4} \alpha_{jk} = 0$.

(ii) Linear mapping of the four-dimensional parameter $\boldsymbol{\alpha}_j$ to the unconstrained three-dimensional real parameter $\boldsymbol{\beta}_j \in \mathbb{R}^3$:

$$\boldsymbol{\alpha}_j = H\boldsymbol{\beta}_j,$$

where $\boldsymbol{\beta}_j = (\beta_{j1},\ \beta_{j2},\ \beta_{j3})^T$, and H is a 4-by-3 matrix with (j,k)-th entry

$$h_{jk} = \begin{cases} 0, & \text{if } j < k \\ d_k, & \text{if } j = k \\ -d_k/(4-k), & \text{if } j > k \end{cases}$$

for $j = 1, \ldots, 4$ and $k = 1, \ldots, 3$. Here $d_1 = 1$ and $d_k = d_{k-1}\sqrt{1 - 1/(4-k+1)^2}$ for $k = 2\ , 3$. This choice of H is symmetric in the sense that $\boldsymbol{\pi} = (0.25,\ 0.25,\ 0.25,\ 0.25)$ is mapped to $\boldsymbol{\beta} = (0,\ 0,\ 0)$.

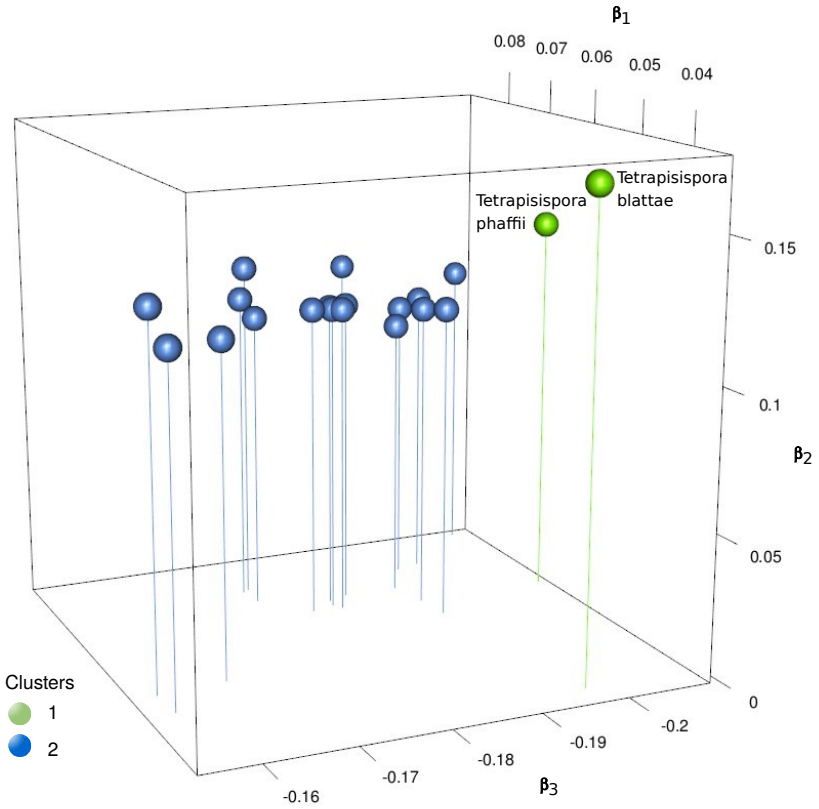


Figure 5.6: Graphical visualisation of the empirical composition of nucleotides for the yeasts data set. Each circle represents a three-dimensional vector $\boldsymbol{\beta}_j$ obtained by transforming the empirical composition $\boldsymbol{\pi}_j$ of species j into $\mathbb{R}^3$. Green and blue colours represent clustering of the $\boldsymbol{\beta}_j$ into two groups according to the k-means clustering procedure with $k = 2$. The non-stationary models place the root on a pendant edge leading to *Tetrapisispora blattae* (cluster 1).

Since the non-stationary model assumes two composition vectors, we divide the parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, n$, into two clusters. This is to check whether the difference in composition of nucleotides between species could explain our rooting results. In order to do this we apply a k-means clustering procedure to the transformed composition, $\boldsymbol{\beta}_j$. The

k-means clustering partitions the data into k clusters such that the sum of squares from the data to the assigned cluster means is minimised. We take $k = 2$ in order to partition the summary statistics $\boldsymbol{\beta}_j$ into two clusters. Figure 5.6 shows the $\boldsymbol{\beta}_j$ plotted in a three-dimensional space, clustered into two groups according to the k-means clustering. Cluster 1 comprises *Tetrapisispora blattae* and *Tetrapisispora phaffii* which appear to be sister taxa on the tree representing the current biological opinion about the palaeopolyploid yeasts (Figure 5.1). We note, that on the trees inferred with the NR and the CAT-GTR models both taxa are located on rather long branches to allow the composition to evolve and become different from the other species (Figures 5.4 and 5.5). The root of the tree inferred with the non-stationary models is located on a pendant edge leading to *Tetrapisispora blattae* (Figure 5.7). We note that even though the species *Tetrapisispora blattae* and *Tetrapisispora phaffii* are clustered together by the k-means clustering procedure (Figure 5.6), the $\boldsymbol{\beta}$ vector of the *Tetrapisispora blattae* is further from the mean of the other cluster.

Thus the rooting on a pendant edge leading to *Tetrapisispora blattae* can be explained

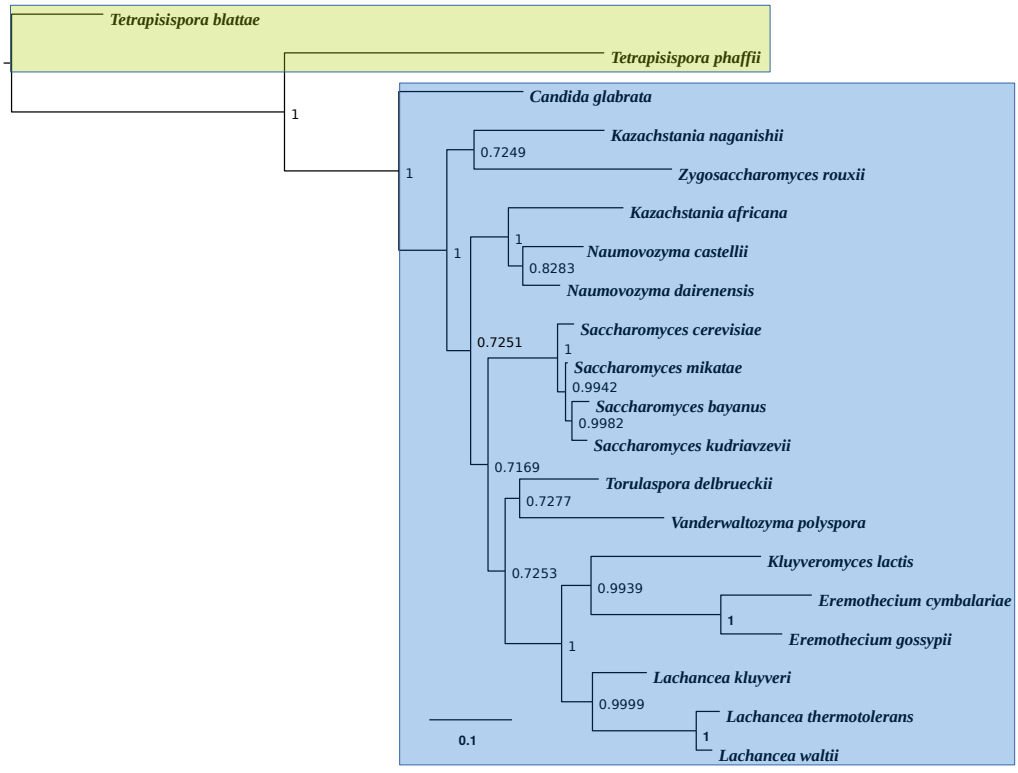


Figure 5.7: Rooted majority rule consensus tree of the palaeopolyploid yeasts data set, inferred under the NRNS model using the Yule prior. The colours represent the two clusters of the summary statistics $\boldsymbol{\beta}$ which were obtained by transforming the empirical composition of the nucleotides into three real numbers (Figure 5.6). The tree inferred with the structured uniform prior looks very similar and so is not shown.

by the fact that the NS model tries to separate two compositionally different groups of species. This suggests that for some data sets modelling non-stationarity with two composition vectors (one vector at the root of the tree and the other as the limiting distribution for the rest of the tree) can be misleading. Compositional heterogeneity in this data set would be better accommodated by a more flexible model, which, for example, allowed more frequent changes in the theoretical stationary distribution, as in the models proposed by Foster (2004); Blanquart & Lartillot (2006); Heaps *et al.* (2014). However, this can lead to computational difficulties in model fitting due to the increase in the complexity of the model.

In order to investigate the composition further we also constructed a tree based on hierarchical cluster analysis of the parameters $\boldsymbol{\beta}$. First we obtained the matrix of euclidean distances between the parameters $\boldsymbol{\beta}$. We then perform a hierarchical cluster analysis which works by assigning each parameter its own cluster and then joining the two most similar clusters iteratively, until there is just a single cluster. The similarity between the clusters is based on the matrix of euclidean distances. The tree constructed according to the hierarchical clustering analysis shows that the *Tetrapisispora blattae* is located at the base of the tree (Figure 5.8). This provides further explanation for the placement of the root on the pendant edge leading to *Tetrapisispora blattae* by the non-stationary models.

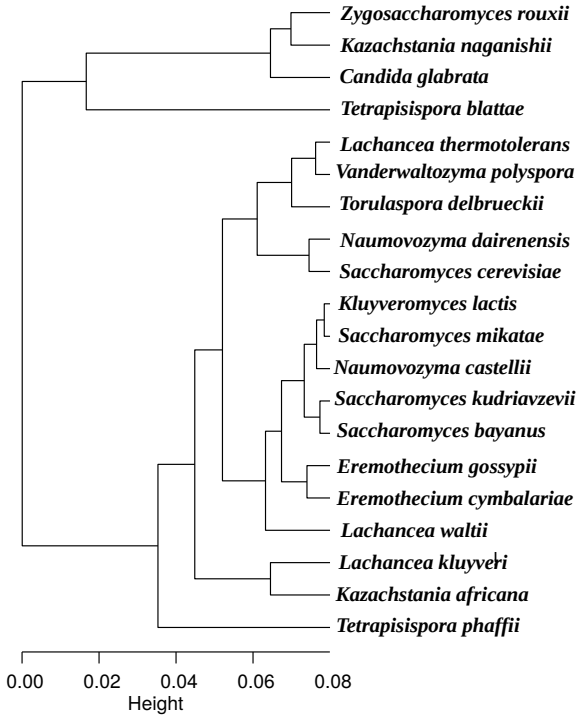


Figure 5.8: The yeasts data set tree based on the hierarchical cluster analysis of the summary statistics $\boldsymbol{\beta}$ which were obtained by transforming the empirical composition of the nucleotides to three-dimensional space (Figure 5.6). The clustering is based on the matrix of euclidean distances.

5.1.3 Posterior predictive simulations

In order to examine the fit between the model and the data, we compared the empirical GC content (proportion of G and proportion of C) for each species with their posterior predictive distributions. To do this, for each iteration of the MCMC algorithm we simulated an alignment using the parameters obtained in this iteration. We then calculated the composition of the nucleotides for these alignments and hence the GC content, thus obtaining the posterior predictive distribution of the empirical GC content. We performed these simulations for each one of the models: NR, NS and NRNS (Figure 5.9).

As expected, the NR model does not account for the heterogeneity in composition. Both NS and NRNS models have posterior predictive mean for the empirical GC content close to the empirical GC content only for species 16 which corresponds to *Tetrapisispora blattae*. In fact, the posterior predictive means for the GC content for all other species are different from that of *Tetrapisispora blattae* (around 0.478 for *Tetrapisispora blattae* with both NS and NRNS models, and around 0.468 (NS model) and 0.46 (NRNS model) for all the other species). This can be explained by the fact that the model places the root on the pendant edge leading to *Tetrapisispora blattae*, so the composition of *Tetrapisispora blattae* is close to the composition at the root, and different from that of all the other species. This provides further evidence that modelling non-stationarity with two composition vectors is not sufficient to describe the compositional heterogeneity in this data set. However, the unrooted topology obtained with the NRNS model is the same as the unrooted topology obtained with the NR model.

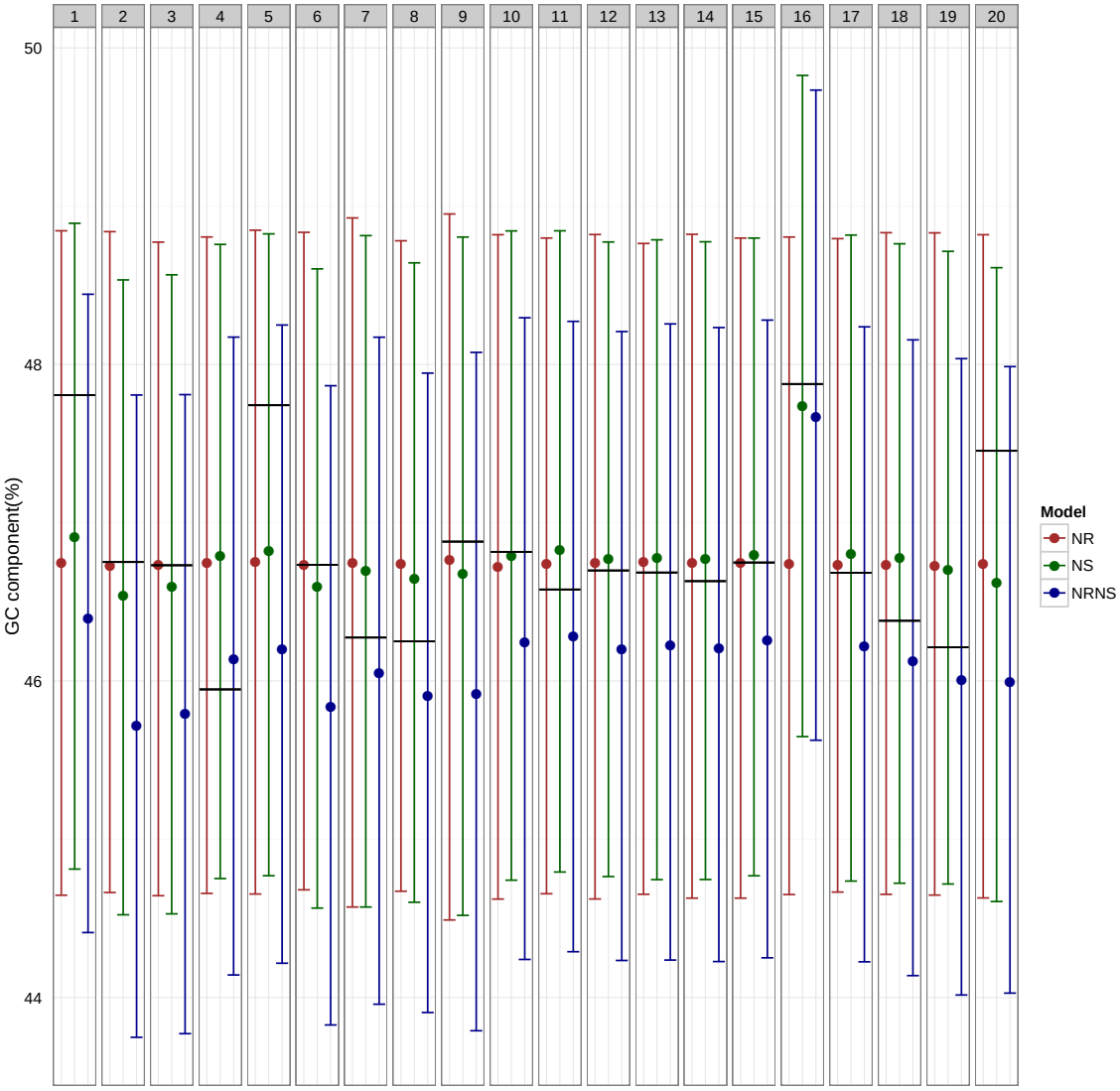


Figure 5.9: Posterior predictive means and 95% credible intervals for the empirical GC content of the yeasts data set. Empirical values are indicated with a horizontal line in each panel. Numbers 1 - 20 correspond to the twenty species of yeasts in the data set:

1 - *Candida glabrata*
2 - *Eremothecium cymbalariae*
3 - *Eremothecium gossypii*
4 - *Kazachstania africana*
5 - *Kazachstania naganishii*
6 - *Kluyveromyces lactis*
7 - *Lachancea kluyveri*
8 - *Lachancea thermotolerans*
9 - *Lachancea waltii*
10 - *Naumovozyma castellii*
11 - *Naumovozyma dairenensis*
12 - *Saccharomyces bayanus*
13 - *Saccharomyces cerevisiae*
14 - *Saccharomyces kudriavzevii*
15 - *Saccharomyces mikatae*
16 - *Tetrapisispora blattae*
17 - *Tetrapisispora phaffii*
18 - *Torulaspora delbrueckii*
19 - *Vanderwaltozyma polyspora*
20 - *Zygosaccharomyces rouxii*

5.2 Rooting the primates data set

5.2.1 12 species data set

As a second example we analysed a set of primates data provided as part of the MrBayes software (Huelsenbeck & Ronquist, 2001). This is a combined nucleic acid data set from subunits 4 and 5 NADH (Nicotinamide Adenine Dinucleotide) dehydrogenase genes. The primates data set has been analysed previously in a maximum likelihood framework. Non-reversible and non-stationary models were fitted to rooted trees with a fixed unrooted topology, and the likelihood values of the rooted trees were then compared. It has been found that non-stationary models were able to infer the root, while non-reversible models were not (Yap & Speed, 2005). Current biological opinion about this data set is summarised in Figure 5.10 (Purvis, 1995; Perelman *et al.*, 2011), with roots 1, 2 and 3 being biologically plausible (located near *Tarsius* and *Lemur*). We analysed this data set with the NR, NS and NRNS models. Our results are consistent with the published analysis: the non-stationary models infer biologically plausible root splits, while with the stationary

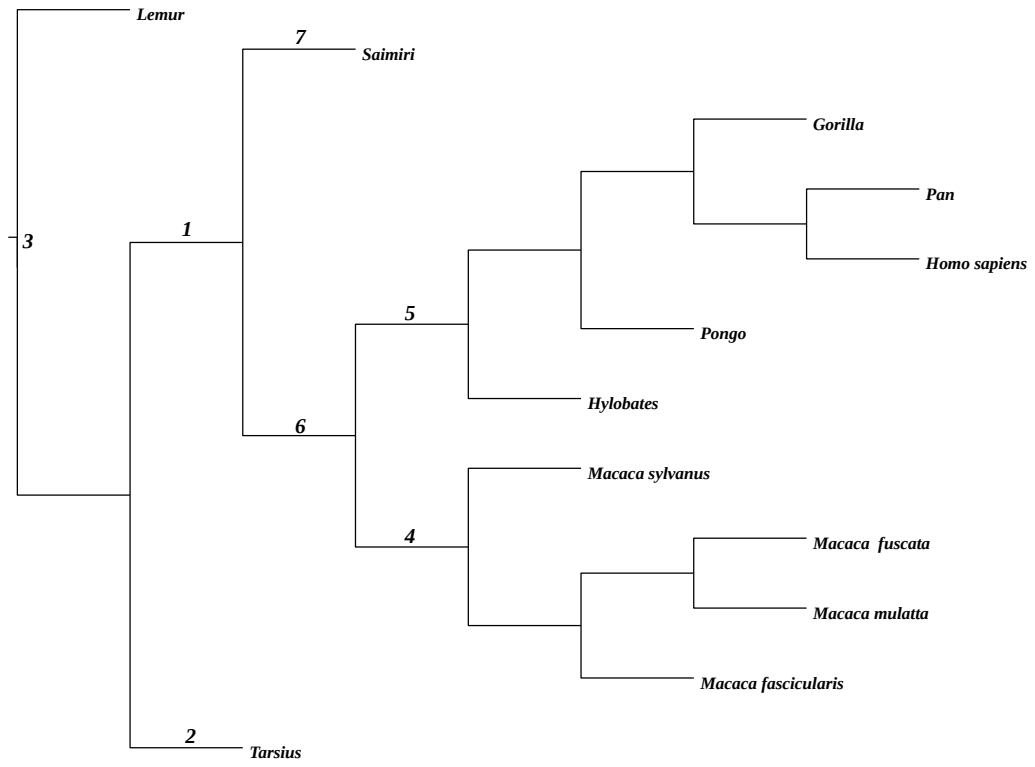
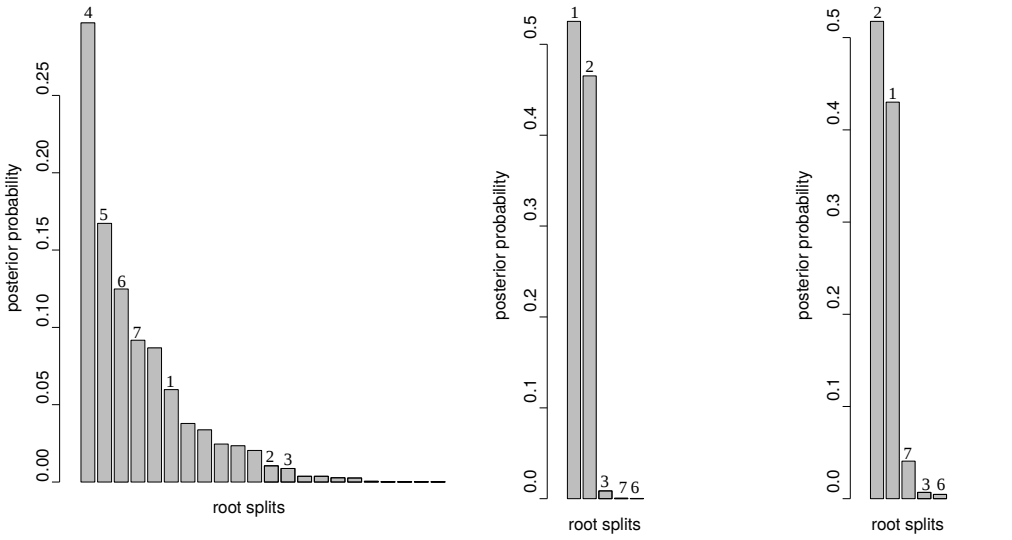


Figure 5.10: Schematic tree of the primates 12 species data set. Numbers 1 - 7 represent root splits obtained in the analyses with our non-reversible and non-stationary models. Biologically plausible roots are roots 1 - 3.

model these root splits have very low support (Figure 5.11). Both NS and NRNS models assign most posterior support to root 2 (on the *Tarsius* branch) and root 1 (on the branch connecting *Tarsius* and *Lemur* to the other species). The other root splits supported by non-stationary models are located on the *Lemur* branch (root 3), *Saimiri* branch (root 7) and on the branch connecting *Tarsius*, *Lemur* and *Saimiri* to the other species (root 6). Thus all the root splits inferred with the non-stationary models are in the vicinity of the biologically plausible root. By contrast, the NR model assigns the highest posterior probability to the roots on the branch leading to the *Macaca* clade (root 4) and on the branch leading to the apes (root 5). These two roots contradict biological opinion about the phylogeny of primates. In terms of the unrooted topology, all three models recovered the widely agreed topology (Figure 5.10) with posterior probability of almost 1.



(a) NR model (b) NRNS model (c) NS model

Figure 5.11: Posterior distribution of the root splits for the primates 12 species data set analysed with (a) NR model, (b) NRNS model, (c) NS model. NR model: the most plausible root split is between the *Macaca* clade and the other species (root 4), the second most plausible root split is between the apes and the other species (root 5); both contradict biological opinion. NS and NRNS models: high posterior probability of the root being somewhere near *Tarsius* and *Lemur* (roots 1 and 2); this is in accord with biological opinion. The roots are mapped in Figure 5.10.

Similarly to the yeasts data set, we transformed the vectors of the empirical composition of the nucleotides $\boldsymbol{\pi}_j$ to unconstrained real vectors $\boldsymbol{\beta}_j \in \mathbb{R}^3$ and applied the k-means clustering procedure with $k = 2$ in order to partition the $\boldsymbol{\beta}_j$ into two clusters. Cluster 1 comprises four species, three of them are in the vicinity of the biologically plausible root (*Tarsius*, *Lemur* and *Saimiri*) (Figure 5.12). We also constructed a tree from the

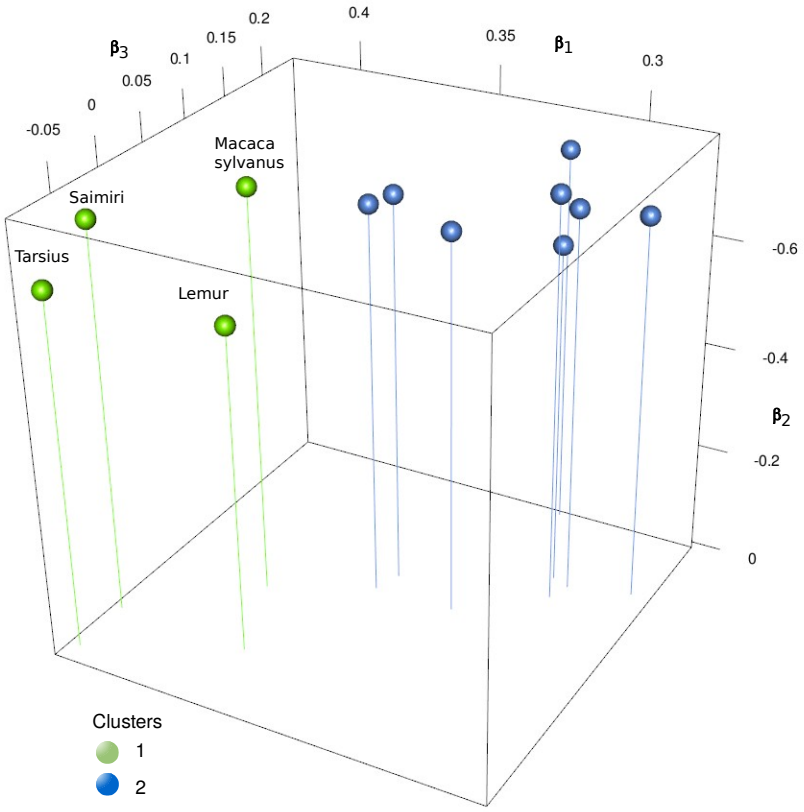


Figure 5.12: Graphical visualisation of the empirical composition of nucleotides for the primates data set. Each circle represents a three-dimensional vector $\boldsymbol{\beta}_j$ obtained by transforming the empirical composition $\boldsymbol{\pi}_j$ of species j into $\mathbb{R}^3$. Green and blue colours represent clustering of the $\boldsymbol{\beta}_j$ into two groups according to the k-means clustering procedure with $k = 2$. *Tarsius*, *Lemur* and *Saimiri* (cluster 1) are the species in the vicinity of the biologically plausible root. Non-stationary models support root positions near *Tarsius* and *Lemur*.

hierarchical cluster analysis. On this tree, *Lemur*, *Tarsius* and *Saimiri* appear as a cluster on one side of the root, which provides additional explanation of the placement of the root around *Tarsius* and *Lemur* by non-stationary models (Figure 5.13).

Analysis of the posterior predictive distributions for empirical GC content shows a rather poor fit for all three models (Figure 5.14). Nevertheless, with the non-stationary models, species 4 and 12 which correspond to *Lemur* and *Tarsius*, respectively, have posterior predictive means for the empirical GC content different from the other species.

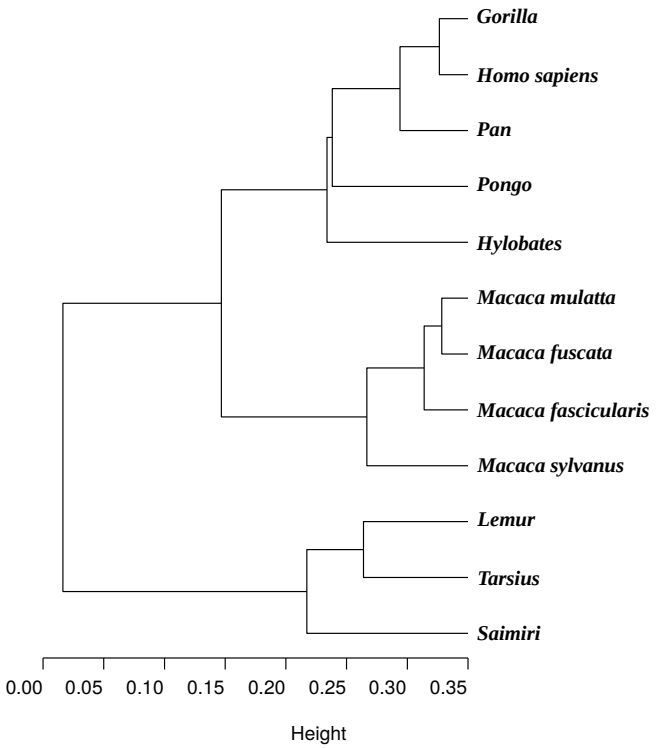


Figure 5.13: The primates data set tree based on hierarchical cluster analysis of the summary statistics $\boldsymbol{\beta}$ obtained by transforming the empirical composition of the nucleotides to three-dimensional space (Figure 5.12). The three species comprising the bottom cluster (*Tarsius*, *Lemur* and *Saimiri*) are in the vicinity of biologically plausible root.

This can be explained by the fact that these species are close to the inferred root position.

The success of the non-stationary model to infer the root in the primates data set can be explained by the high level of non-stationarity inferred from the data. We define the *level of non-stationarity* as an Euclidean distance between the inferred composition at the root and the inferred theoretical stationary distribution. The posterior mean for the level of non-stationarity in the primates data set is 0.279 which is even higher than the "high" level of non-stationarity in the simulated data (0.203). In the yeasts data set, the posterior mean for the level of non-stationarity is 0.146 which is close to the level of "moderate" non-stationarity (0.1) in the simulations of the NS model. For the NRNS model, the posterior mean for the level of non-stationarity is also higher in the primates data set in comparison to the yeasts data set (0.36 vs. 0.279). However, unlike the yeasts data set, the posterior for σ in the primates dataset offers high support for values of σ around zero (Figure 5.3). This can explain the failure of the NR model to infer the root in the primates data set.

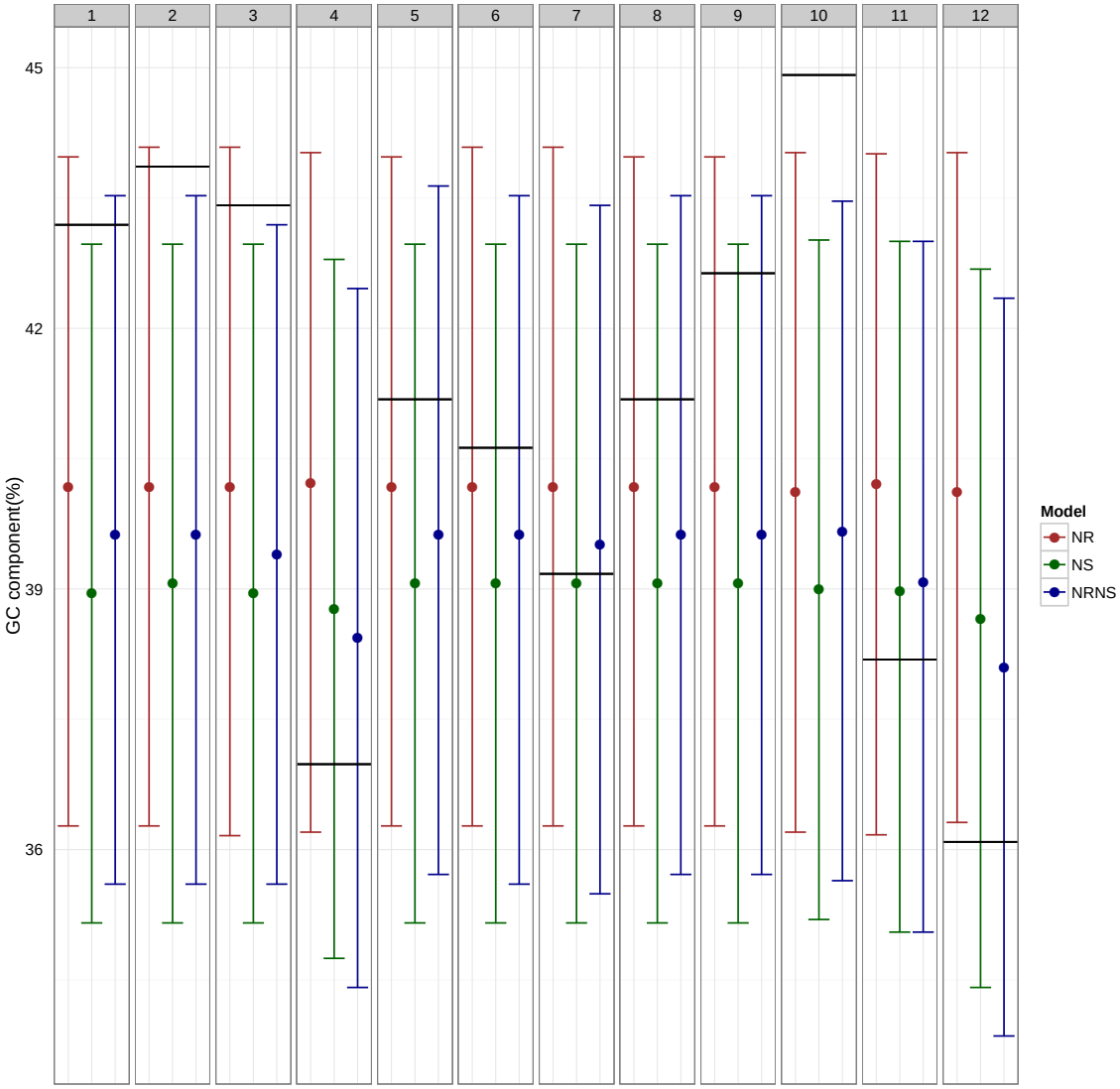


Figure 5.14: Posterior predictive means and 95% credible intervals for the empirical GC content of the primates 12 species data set. Empirical values are indicated with a horizontal line in each panel. Numbers 1 - 12 correspond to the twelve species of primates in the data set:

1 - *Gorilla*
2 - *Homo sapiens*
3 - *Hylobates*
4 - *Lemur*
5 - *Macaca fascicularis*
6 - *Macaca mulatta*
7 - *Macaca sylvanus*
8 - *Macaca fuscata*
9 - *Pan*
10 - *Pongo*
11 - *Saimiri*
12 - *Tarsius*

5.2.2 Expanded primates data set

In order to demonstrate scalability of the models, we have expanded the above data set to include most of the species analysed in Perelman *et al.* (2011). The expanded data set comprises 38 species (the schematic rooted tree is represented in Figure 5.15).

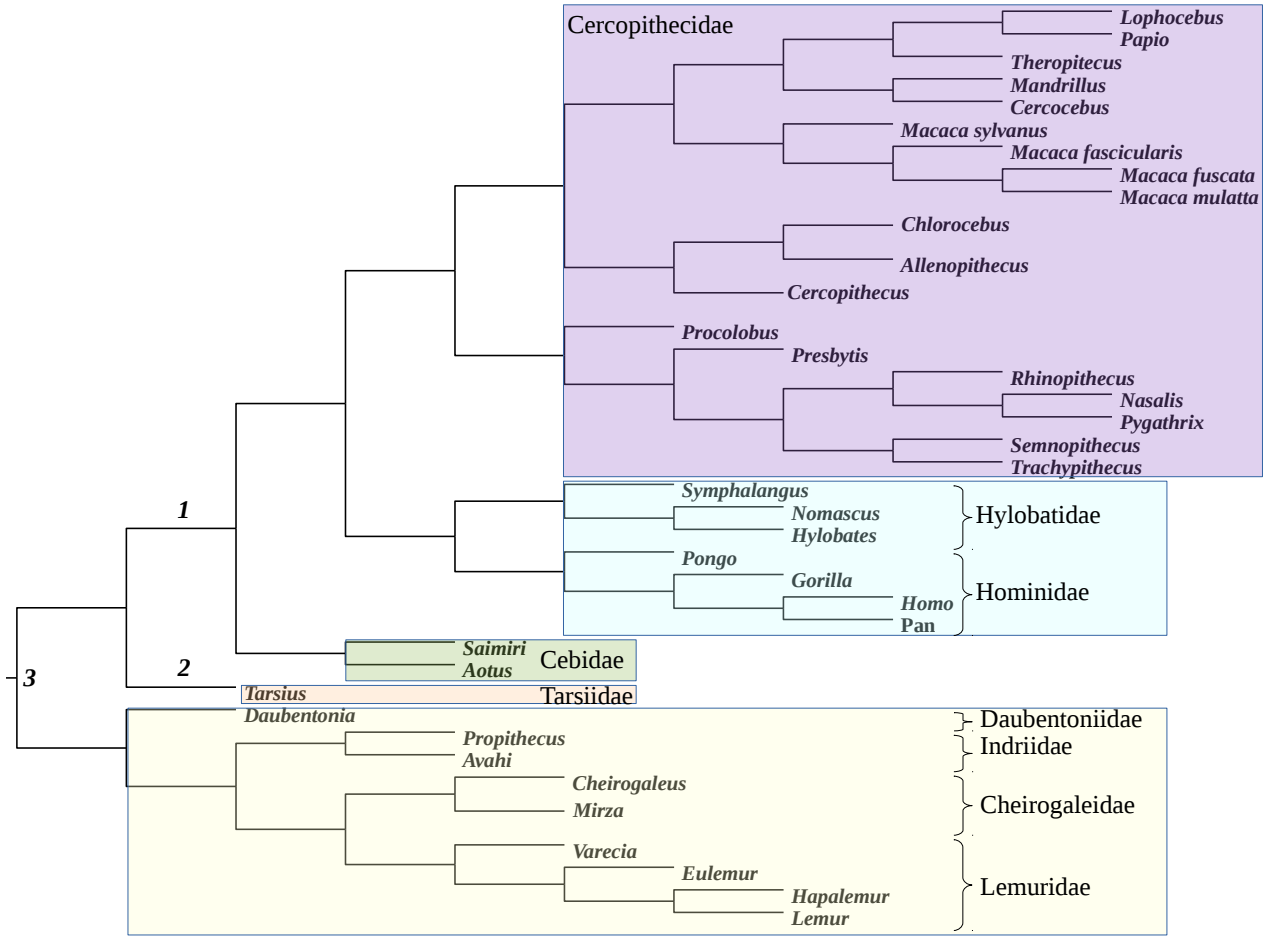
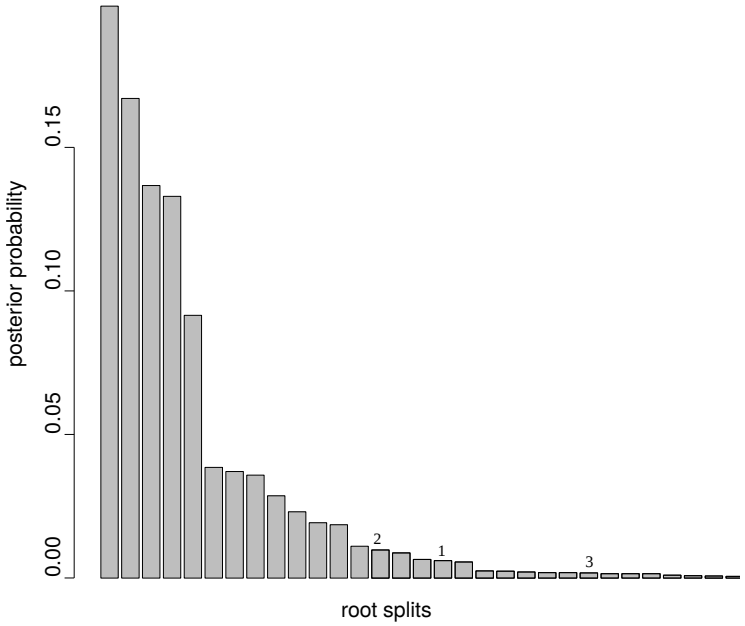
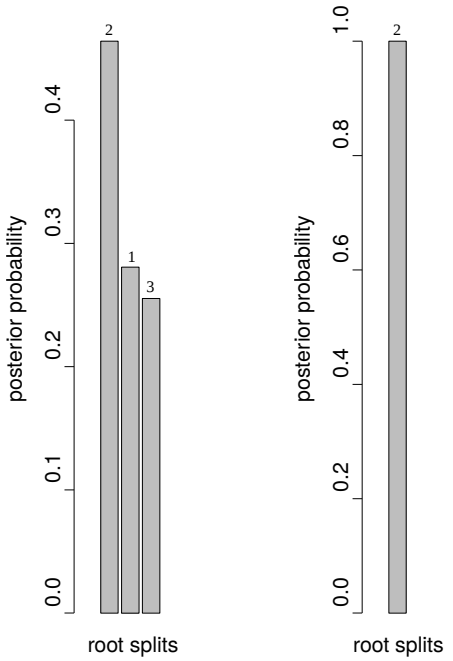


Figure 5.15: Schematic rooted tree of the expanded primates data set, comprising 38 species. Branches 1 - 3 represent the region of biologically plausible root positions.

We analysed this data set with the NR, NS and NRNS models. The results are consistent with those from the smaller data set. The NR model recovers the biologically plausible root positions with very low posterior probability. On the other hand, non-stationary models support the root splits at the region of the true root: the NS model infers root 2 with posterior probability 1; the NRNS model supports three different root splits: root 2 (posterior probability = 0.46), root 1 (posterior probability = 0.28) and root 3 (posterior probability = 0.26) (Figures 5.15 and 5.16).



(a) NR model



(b) NRNS model

(c) NS model

Figure 5.16: Posterior distribution of the root splits for the primates 38 species data set, inferred with the NR, NRNS and NS models: (a) the NR model recovers the biologically plausible roots (roots 1 - 3) with low posterior support; (b) the NRNS model supports three biologically plausible roots (roots 1 - 3); (c) the NS model supports one of the biologically plausible roots (root 2). The roots are mapped in Figure 5.15.

The consensus tree recovers the major relationships amongst primates but for the placement of *Ceirogaleus* which is clustered with *Cercopithecidae* in all of our analyses (Figure 5.17) (the unrooted topology of the consensus trees inferred with the NR, NS and NRNS models are very similar). While the peculiar placement of *Ceirogaleus* requires further investigation, we note that this placement is consistent with the analysis under the CAT-GTR model (Figure 5.18).

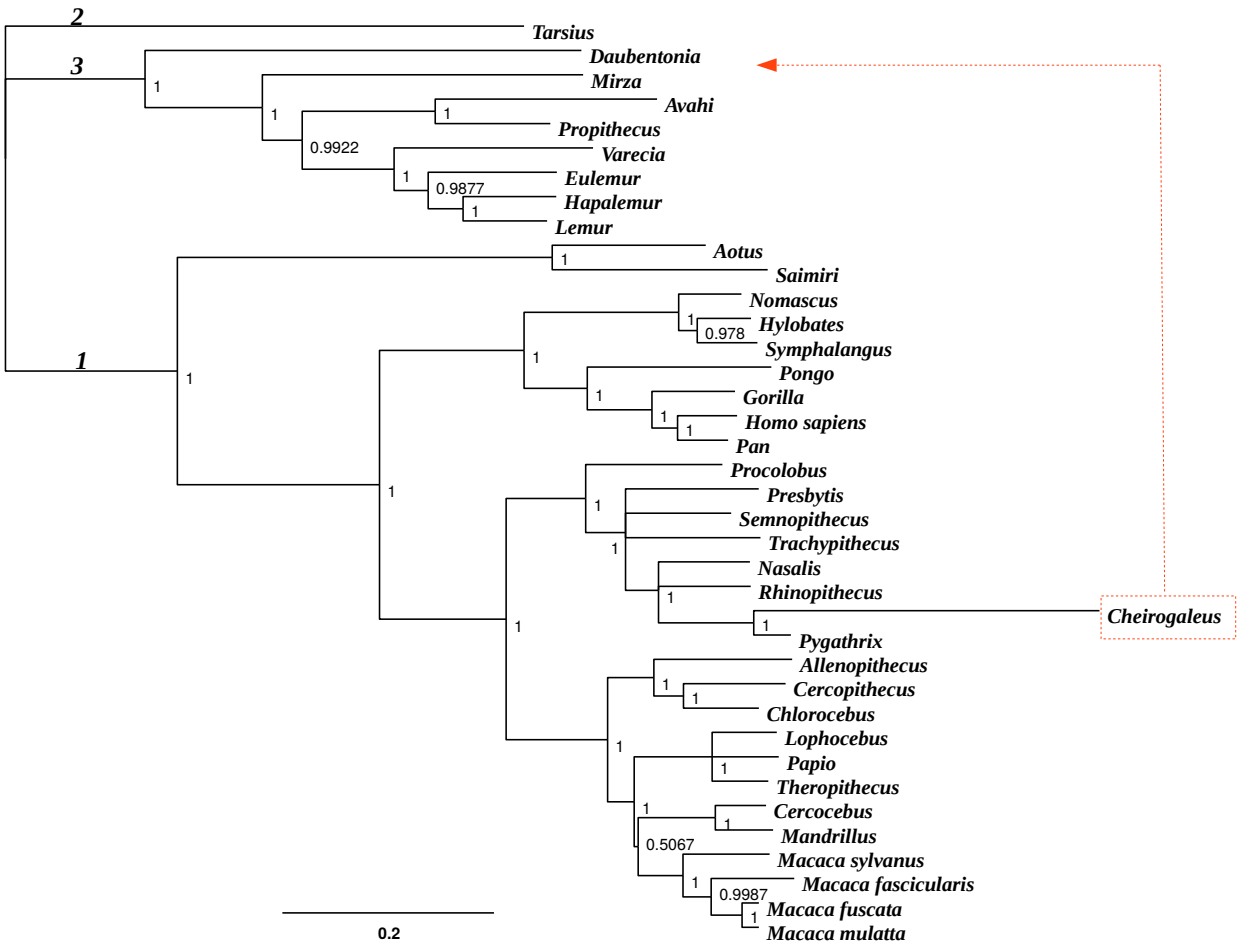


Figure 5.17: Rooted consensus tree of the primates 38 species data set, inferred with the NRNS model. The unrooted topology corresponds to that of the schematic tree (Figure 5.15) but for the placement of the *Ceirogaleus*. Roots 1 - 3 correspond to the roots mapped on the schematic tree (Figure 5.15).

Interestingly, the plot of the posterior density for σ is suggestive of a large degree of non-reversibility in the data (Figure 5.3). However, the non-reversible model infers the biologically plausible root with very low posterior probability. This can be explained by the fact that the data contain a very high level of non-stationarity (posterior mean for the level of non-stationarity is 0.35, which is bigger than the "high" level of non-stationarity in

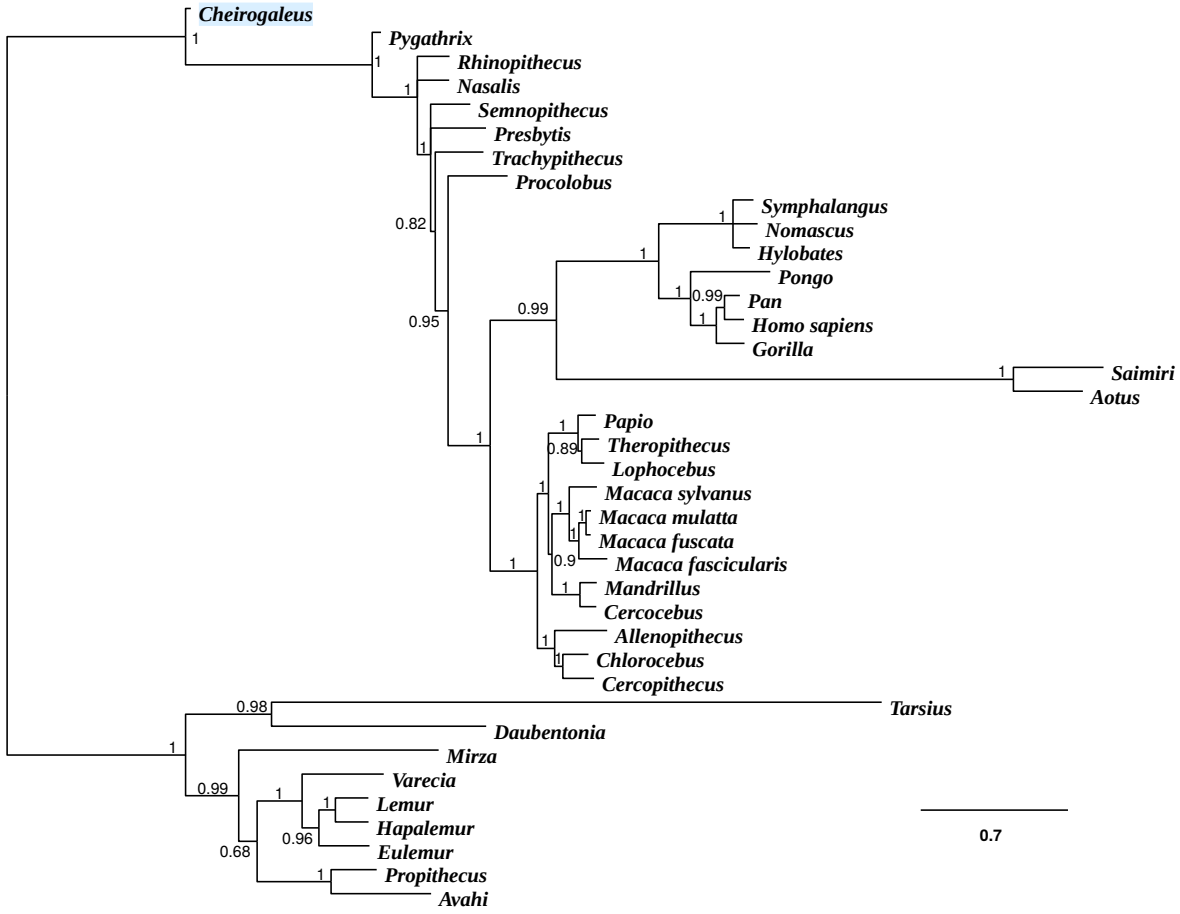


Figure 5.18: Unrooted consensus tree of the primates 38 species data set, inferred with the CAT-GTR. The *Ceirogaleus* (shaded in blue) is placed within the *Cercopithecidae*. This placement is consistent with the results of our non-reversible and non-stationary models (Figure 5.17); however, it contradicts the placement of *Ceirogaleus* on the schematic tree (Figure 5.15).

the simulated data), and we have previously shown in the simulations for the NRNS model that the effect of non-stationarity seems to dominate over the effect of non-reversibility (see Section 4.2.2).

5.3 Rooting the tree of life

Finally we applied the models to a data set for which there is still debate about the unrooted topology and root position: the ribosomal tree of life. The debates are centred on two hypotheses. According to the three domains hypothesis, the Archaea are monophyletic, sharing a common ancestor with the Eukaryota (Woese, 1990). The other hypothesis, called the eocyte hypothesis, suggests that the Archaea are paraphyletic and

the Eukaryota originated from within the Archaea (Lake, 1988; Rivera & Lake, 1992; Cox *et al.*, 2008). Recent analyses of the tree of life ribosomal RNA data have demonstrated that inference on the tree is sensitive to the substitution model that is fitted. When homogeneous (and therefore stationary) models are used for the analysis they often recover the three domains tree, while heterogeneous models generally recover the eocyte tree (Cox *et al.*, 2008; Williams *et al.*, 2012; Heaps *et al.*, 2014). In addition to heterogeneous models, there is also external evidence for the eocyte hypothesis. For example, newly discovered archaeal species whose genomes encode many eukaryote-specific features, provide additional support for the eocyte hypothesis. (Spang *et al.*, 2015).

5.3.1 16 species data set

We analysed a previously published 16-species concatenated rRNA alignment containing 761 sites from small subunit ribosomal RNA (Heaps *et al.*, 2014). The data set comprises archaebacterial, bacterial and eukaryotic species, including the recently discovered archaeal groups: Thaumarchaeota, Aigarchaeota and Korarchaeota. These new groups are closely related to Crenarchaeota and together they form the so-called TACK superphylum (Guy & Ettema, 2011; Kelly *et al.*, 2011; Williams *et al.*, 2012; Lasek-Nesselquist & Gogarten, 2013). The analysis with the NR model recovered a widely accepted root (i.e. between the Bacteria and Archaea) with posterior probability 0.72 (Figure 5.19). This root is

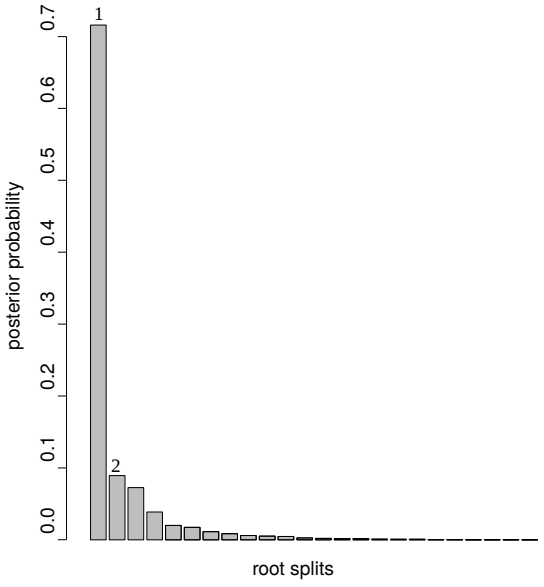


Figure 5.19: The posterior distribution of the root splits of the tree of life 16 species data set for the NR model analysed with the Yule prior. Different bars on the plot represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). The root split on the branch leading to the Bacteria has the highest posterior probability (root 1). Root 2 is placed within the Bacteria (on the branch leading to *Rhodopirellula baltica*). The roots are mapped in Figure 5.20.

supported by paralogue rooting methods (Iwabe *et al.*, 1989; Gogarten *et al.*, 1989; Brown & Doolittle, 1995) and analysis of genome networks (Dagan *et al.*, 2010). The plot of the posterior density for σ shows evidence of a substantial amount of non-reversibility in the data (Figure 5.3). In terms of the unrooted topology the NR model recovered the classic three-domains topology, in which the eukaryotes emerge as the sister group to a monophyletic Archaea (Figure 5.20). This result is not surprising given that the three-domains tree has been previously supported by analyses with stationary models (Gouy & Li, 1989).

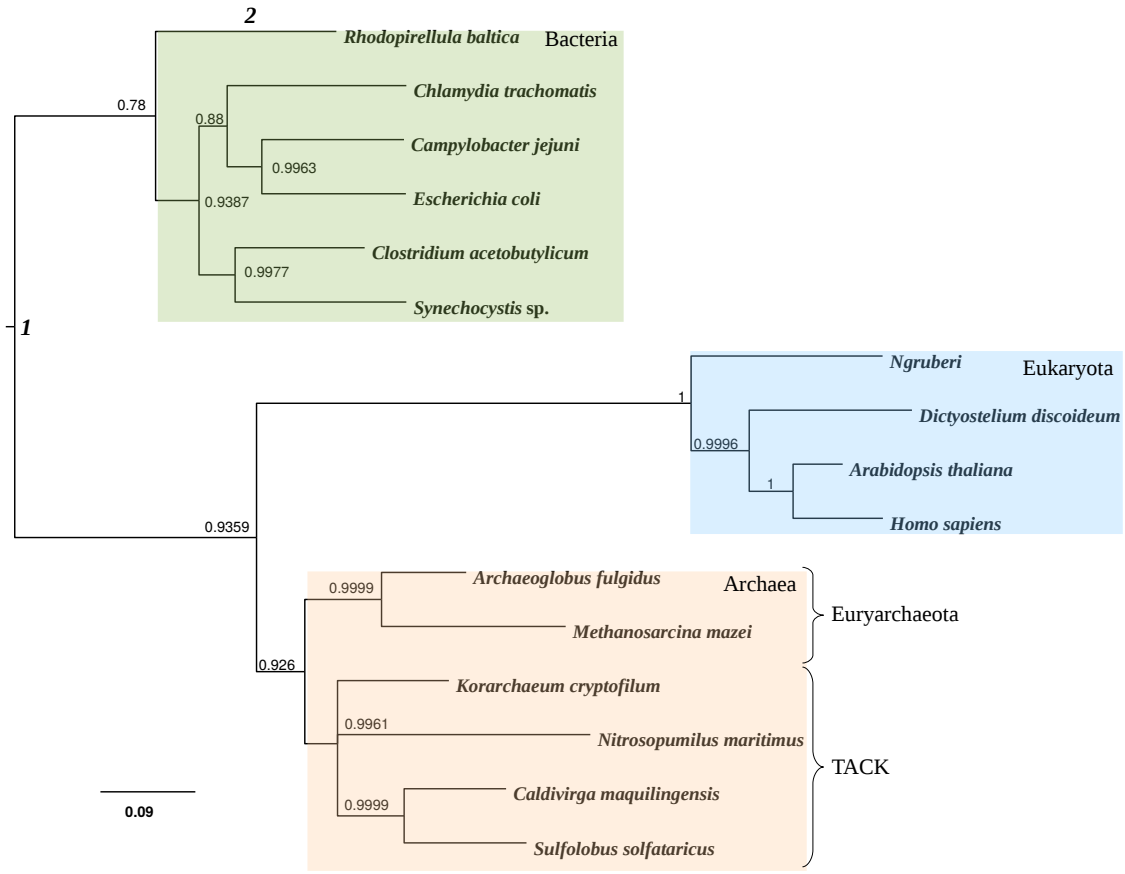


Figure 5.20: Rooted majority rule consensus tree of the tree of life 16 species data set inferred with the NR model and the Yule prior. Roots 1 and 2 have the highest and the second highest posterior support respectively; both roots are plausible from a biological point of view.

We also analysed this data set with the non-stationary model (NRNS). The analysis recovered rather unusual relationships between the three domains of life. On the consensus tree inferred with the NRNS model the Archaea is paraphyletic, with the root placed between the TACK superphylum and all the other species on the tree (Figure 5.21). The root split on the consensus tree has the highest posterior support, and the root split with the second highest posterior support is also within the TACK superphylum (on a pendant edge leading to *Caldivirga maquilingensis*, Figure 5.22). This result can be explained by the fact that the archaeal species are very different in composition from the other species, as shown in the analysis of the composition of nucleotides transformed to the three-dimensional space and clustered into two clusters (Figure 5.23). Again, this suggests that there is a need to account for compositional heterogeneity in a more sophisticated way while implementing non-stationary analyses.

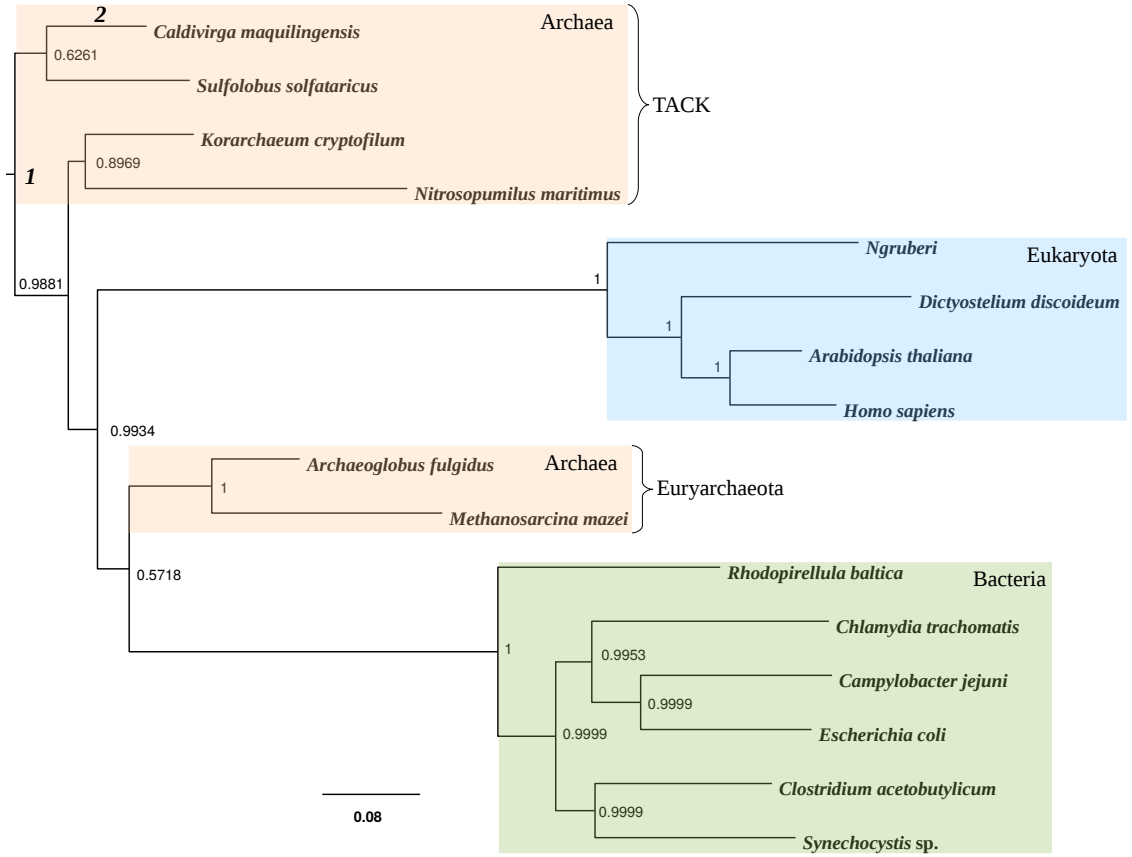


Figure 5.21: Rooted majority rule consensus tree of the tree of life 16 species data set inferred with the NRNS model and the Yule prior. Roots 1 and 2 have the highest posterior probability in our analysis. However, the support for these roots has not been reported previously.

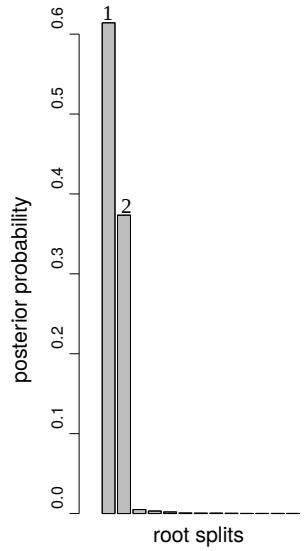


Figure 5.22: The posterior distribution of the root splits of the tree of life 16 species data set for the NRNS model analysed with the Yule prior. Different bars on the plot represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). The root split on the branch separating the TACK superphylum from the other species has the highest posterior probability (root 1). Root 2 is placed within the TACK superphylum (on the branch leading to *Caldivirga maquilingensis*). The roots are mapped in Figure 5.21.

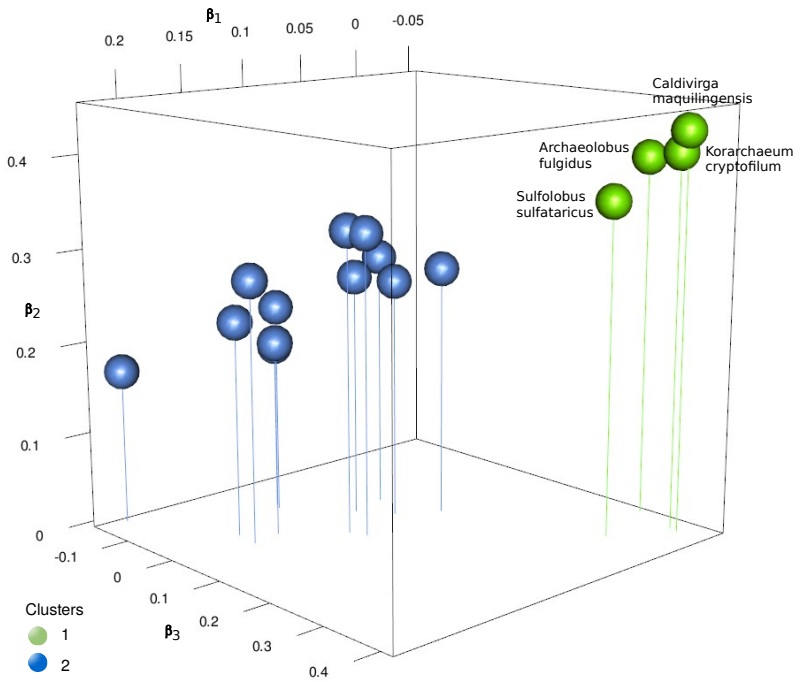


Figure 5.23: Graphical visualisation of the empirical composition of nucleotides for the the tree of life 16 species data set. Each circle represents a three-dimensional vector $\boldsymbol{\beta}_j$ obtained by transforming the empirical composition $\boldsymbol{\pi}_j$ of species j into $\mathbb{R}^3$. Green and blue colours represent clustering of the $\boldsymbol{\beta}_j$ into two groups according to the k-means clustering procedure with $k = 2$. The posterior modal root inferred with the NRNS model separates *Caldivirga maquilingensis* and *Sulfolobus solfataricus* (cluster 1) from the rest of the species.

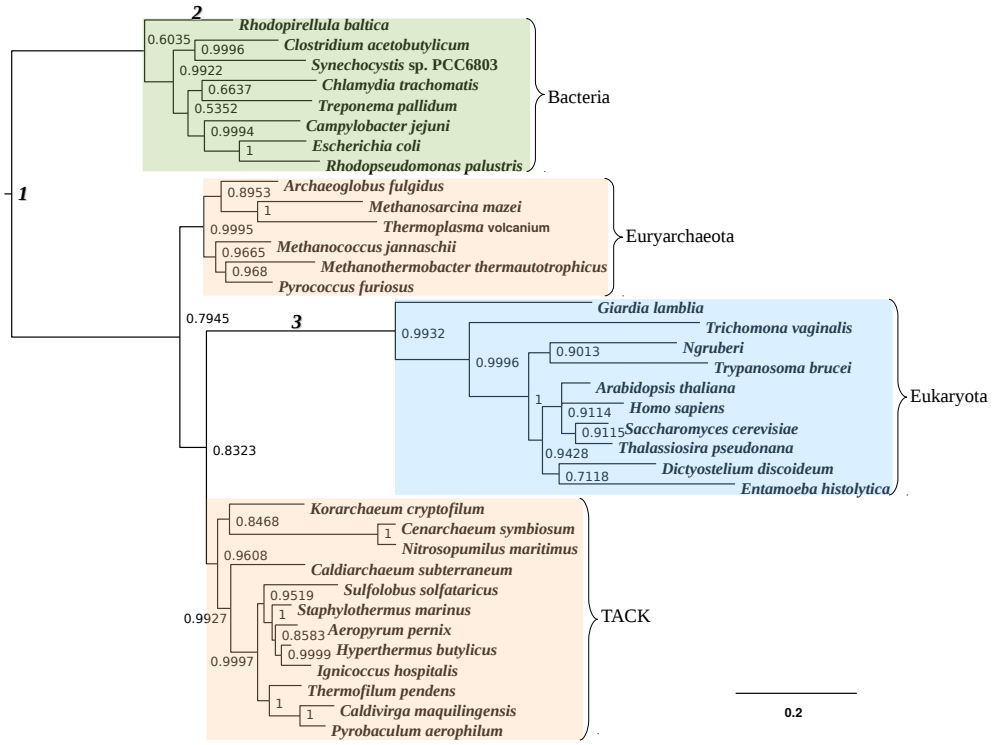
5.3.2 36 species data set

Here we analysed aligned concatenated large and small subunit ribosomal RNA sequences from the archaebacterial, bacterial and eukaryotic species comprising 36 taxa (1734 sequence positions). Previous analyses of this data set recovered an eocyte topology, however these analyses were not able to infer the root because they used stationary substitution models based on reversible rate matrices (Williams *et al.*, 2012). We also analysed these data with both the NR and NR2 models using both the Yule prior and the structured uniform prior (the analyses with non-stationary models did not converge). All the analyses recovered the eocyte topology with similar posterior support. Figure 5.24 shows the majority rule consensus tree from the analysis with the Yule prior and with the structured uniform prior.

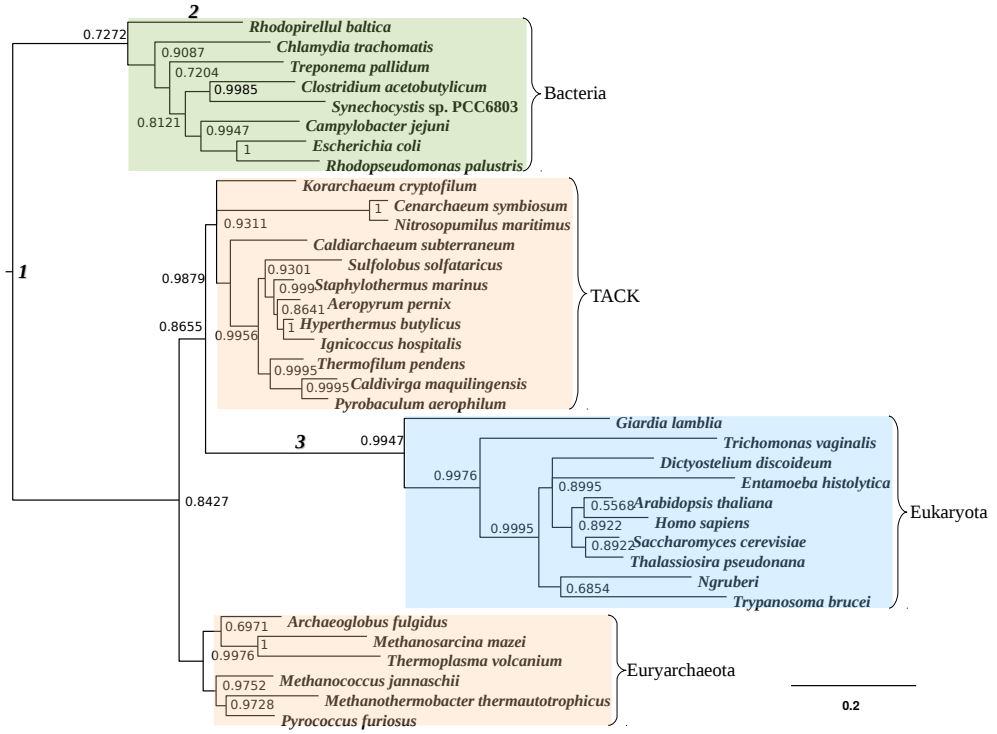
Figure 5.25a shows that the analysis with the Yule prior assigned high posterior support to two roots splits - one on the branch leading to Bacteria (root 1 in Figure 5.24), the other within the Bacteria, on the branch leading to *Rhodopirellula baltica* (root 2 in Figure 5.24). This inference is in accord with current biological opinion about the root of the tree of life, which places the root either on the branch leading to the Bacteria, or within the Bacteria (Baldauf, 1996; Cavalier-Smith, 2006; Skophammer *et al.*, 2007). However, in the analysis performed with the structured uniform prior, the support for the root within the Bacteria decreased and that for the the root on the bacterial branch increased (Figure 5.25b). This analysis illustrates the sensitivity of the inference to the choice of topological prior, and confirms the importance of prior choice in Bayesian phylogenetics.

Summary

In this chapter we have analysed five experimental data sets with non-reversible and non-stationary models. In the analysis of the yeasts data set we have found that non-reversible models are able to extract useful information about the root of the tree, while non-stationary models turned out to be misleading due to presence of some species with a very different composition of nucleotides to other species. On the other hand, the analysis of the primates data set illustrated the success of non-stationarity models, presumably because of the high level of non-stationarity in the data. Analysis of the tree of life with the non-reversible models supports the widely agreed root, however there is a disagreement between the unrooted topologies inferred from the two data sets of the tree of life. While the 16 species data set recovered the classic three-domains topology, the 36 species data set recovered the eocyte topology, thus confirming the importance of taxonomic sampling for inferring ancient evolutionary relationships. The analysis also highlighted the substantial sensitivity of the root inference to the choice of topological priors.

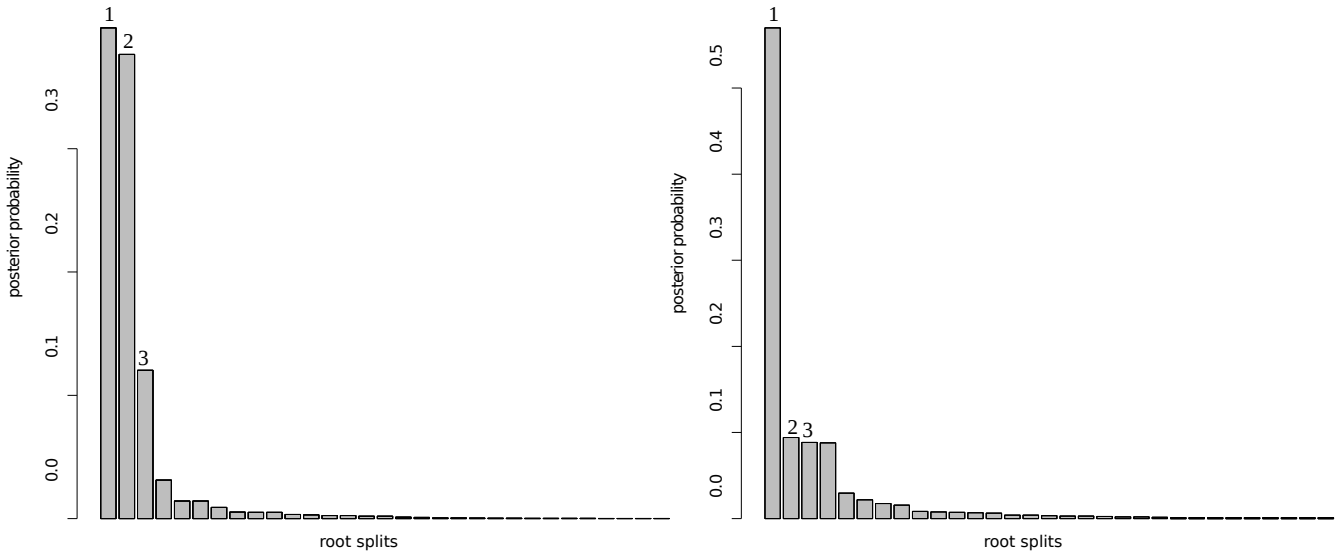


(a) Yule prior



(b) Structured uniform prior

Figure 5.24: Rooted majority rule consensus tree of the tree of life 36 species data set inferred with (a) Yule prior and (b) structured uniform prior. Roots 1 - 3 were inferred in the analysis with our non-reversible models.



(a) Yule prior

(b) Structured uniform prior

Figure 5.25: The posterior distribution of the root splits of the tree of life 36 species data set for the NR model analysed with (a) Yule prior and (b) structured uniform prior. Different bars on the plot represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). The root split on the branch leading to the Bacteria has the highest posterior probability (root 1). Root 2 is placed within the Bacteria (on the branch leading to *Rhodopirellula baltica*) and root 3 is placed on the branch leading to the Eukaryota. The roots are mapped in Figure 5.24.

Chapter 6

Conclusions and future work

This chapter summarises the overall contribution of this thesis to the literature, gives a summary of overall conclusions and discusses potential future work.

6.1 Contributions

Standard phylogenetic models are unable to infer the root of a phylogenetic tree because their likelihood functions do not depend on the root position. Most models that allow root inference are based on relaxing the assumption of homogeneity of the Markov substitution process and are constructed by assigning different reversible rate matrices to different parts of the tree. Whilst biologically persuasive, such non-homogeneous models are, however, highly parameterised, which causes computational challenges in model-fitting. The main contribution of the thesis is in improving understanding of the potential of simpler assumptions which relax the requirement of reversibility and stationarity to enable root inference. By investigating homogeneous non-reversible models which require only one rate matrix we found that non-reversibility is a useful feature of the data for inferring the position of the root (Sections 3.1.5, 3.2.4, 5.1.1). We also highlighted limitations of our approach. The simulation study in Chapter 3 shows that a substantial amount of non-reversibility in the data is needed in order to infer the root correctly. These simulations illustrate that when the level of non-reversibility is low, the posterior probability of the true root is small. Analysis of real data shows that the signal from non-reversibility can be masked by other signals present in the data (Section 5.1.2).

Sensitivity of phylogenetic analysis to the choice of prior for branch lengths and the topology itself has been noted previously (Yang & Rannala, 2005; Alfaro & Holder, 2006). In the thesis we investigated the robustness of posterior root inferences to conflicting prior and likelihood information concerning the rooted topology and branch lengths.

We have contributed to the understanding of the effect of non-stationarity to the root

inference by studying the effect of relaxing the stationarity assumption in a similar manner to how we relaxed the reversibility assumption. In the simulation study we found that the signal of non-stationarity in the data is useful for root inference. However, analysis of real data shows limitations in applying the model to data sets where some species are very different in composition from the others. We also investigated the confounding effects of non-reversibility and non-stationarity and found that, for the simulation cases we studied, the combination of different levels of non-reversibility with non-stationarity does not necessarily improve the root inference (Section 4.2.2). In the analysis of experimental data we found that the signal from non-stationarity is stronger and hence has the potential to mask the signal from non-reversibility (Section 5.1.2).

We also made a methodological contribution by proposing a structured uniform prior for rooted topologies. This prior is an approximation of a biologically defensible Yule prior, with the advantage of being uniform over rooted topologies and less computationally demanding. Overall, our work extends earlier work on non-reversible and non-stationary models which has been limited to fixed unrooted topologies (Huelsenbeck *et al.*, 2002; Yap & Speed, 2005). The principal contribution of our work is that it facilitates inference of both unrooted topology and root position. Although our approach has limitations it represents a direction towards improving root inference of phylogenetic trees.

6.2 Conclusions

In this thesis we presented two substitution models in which changing the root position changes the likelihood of the tree. We started by proposing two hierarchical non-reversible but stationary models in Chapter 3, the NR model and the NR2 model. These models have rate matrices which are centered on that for a (reversible) HKY85 model, but they differ in the structure of the perturbation. The NR model uses one perturbation component which allows a departure from the HKY85 structure. In contrast, the NR2 model utilises two variation components and the perturbation is performed on the space of reversible and non-reversible rate matrices separately. This separation allows us to judge the extent of the different types of perturbation. In simulations for the NR model, we analysed the model with two topological priors (the Yule prior and its approximation, the structured uniform prior) and with five different values of the non-reversibility perturbation parameter $\sigma = 0,\ 0.05,\ 0.1,\ 0.2,\ 0.3$. For each topological prior and each value of the perturbation parameter we analysed nine different alignments, simulated using either the same or different rate matrices. Our simulations show that as expected, for $\sigma = 0$ the data contain no information about the root. However, for $\sigma > 0$ the posterior is often concentrated around the true root, though not in all cases. For larger values of σ the root is inferred more accurately.

We note that departures from the HKY85-structure do not necessarily lead to a non-reversible model. In fact, they could lead to a more general reversible rate matrix. As such the two types of deviation are confounded and so for any given data set, large values of the perturbation parameter do not necessarily provide evidence of non-reversibility. The NR2 model addresses this issue by using a two-stage process to perturb the underlying HKY85 rate matrix. The first perturbation is within the space of GTR matrices, locally perpendicular to the subspace of HKY85 matrices. The second perturbation is within the space of general rate matrices but locally perpendicular to the subspace of GTR matrices. Splitting the perturbation up allowed us to simulate rate matrices with a larger degree of non-reversibility whilst maintaining a biologically plausible stationary distribution.

A similar set of simulations was performed for the NR2 model. The data were simulated with the same value of the reversible perturbation ($\sigma_R = 0.1$) and five different values of the non-reversible perturbation ($\sigma_N = 0,\ 0.1,\ 0.25,\ 0.5,\ 1$). As σ_N increases, the posterior probability of the true root increases, and for larger values of σ_N the model was able to infer the root with very high posterior support for all cases. Under both models the true unrooted topology was inferred with high posterior support.

In the simulation study we also addressed the issue of the sensitivity of root inferences to conflict between the prior and the likelihood. We showed that long branches can potentially mislead the root inference. However, in the absence of very long branches, non-reversible models can extract information from the data about the root even in the face of prior-likelihood conflict (Section 3.1.5).

However, the stationarity assumption of the NR and NR2 models is not realistic from a biological point of view. We therefore investigated relaxing the stationarity assumption of the HKY85 model. Here we employed a similar strategy to that of relaxing the reversibility assumption. In Chapter 4 we presented a non-stationary model in which the composition vector at the root is centered on the stationary composition. Simulation experiments showed that the higher the level of non-stationarity in the data, the better the root inference (Section 4.1.3). We then combined the idea of non-reversibility and non-stationarity in one model which is non-reversible and also has a composition at the root vertex centered on the theoretical stationary distribution. A simulation study showed the potential problem of confounding between the two signals, with the effect of non-stationarity dominating the effect of non-reversibility (Section 4.2.2).

The analysis of experimental data presented in Chapter 5 highlights the success and the limitation of our models. The non-reversible models are able to extract useful rooting information from real biological data. On the other hand, analyses with the non-stationary models suggest that modelling non-stationarity with two composition vectors can be misleading for certain data sets in which the composition of some species is very different from the others. These data sets require more complex modelling of compositional heterogene-

ity, but more complex models are typically more highly parameterised and often difficult to fit to data. Therefore finding a trade-off between model complexity and computational tractability of model-fitting seems desirable.

6.3 Future work

One of the limitations of the NR model is that the perturbation parameter σ is inferred from only 12 off-diagonal elements of the rate matrix. Inferring σ is important because the model assumes that the information about the root comes from the non-reversibility of the data. Therefore it would be interesting to look at a rate matrix of larger dimension. We extended the NR model so that it could be applied to Dayhoff-recoded data. The corresponding rate matrix comprises 30 off-diagonal elements. However, we did not perform an exhaustive simulation study, and also did not apply this model to experimental data. Similarly, the model could be extended to amino acid data. However, in this case, there might be computational challenges owing to the large number of the off-diagonal elements of the rate matrix (380 off-diagonal elements). In particular, this may require a joint proposal and perhaps a parallel implementation of the MCMC algorithm.

One more potential direction of improvement is modelling across-site heterogeneity in a more complex way. Currently we model across-site heterogeneity through linear scaling of the rate matrix where the scaling variable is drawn from a discrete version of a gamma $\mathrm{Ga}(\alpha, \alpha)$ distribution with four categories (Yang, 1994). This way is convenient computationally because the discretisation of a gamma distribution greatly simplifies the calculation of the likelihood of a phylogenetic tree. However, there is no biological necessity for accounting for across-site heterogeneity in such a way, because only the rate of evolution is allowed to change, and not the evolutionary pattern. Ideally a more complex way of modelling across-site heterogeneity which captures main features of the underlying biological process of variation at different sites would make the model more realistic and improve the inference.

Prior information is an important issue in Bayesian phylogenetics. Our analyses show the sensitivity of the models to the topological priors. The analyses also underpin the problem of long internal branches for root inference. Standard phylogenetic priors for branch lengths assign little prior density to long branches and so favour rooting on long branches because it results in dividing a long branch into two short branches. Therefore, a potential direction of development could be to construct a joint prior for rooted phylogenies and branch lengths which, for example, allows for long branches, or alternatively, penalises rooting on long branches.

When applying statistical methods in phylogenetics, integration of the underlying biology is of great importance. Biological information from morphological and fossils data,

for example, could be incorporated into the analysis in order to reduce the posterior uncertainty. Adding different sources of information about the root such as lateral gene transfers (LGTs) and gene duplications and losses could be also of use and value (Abby *et al.*, 2012; Boussau *et al.*, 2013). For example, if it is believed that gene loss is more likely than acquisition of genes (and data suggest that this is the case), then patterns of gene presence or absence in extant genomes will favour particular rooted tree topologies. Combining different sources of information about the root either via a hierarchical Bayesian model or by a sequential Bayesian approach in which various data sources refine our knowledge of the root position, would contribute to future research of rooting phylogenetic trees.

Appendix A

Proof of the existence of the stationary distribution

In order to prove that the stationary distribution exists we (a) prove the existence of the left eigenvector $\boldsymbol{\pi}$ which correspond to the eigenvalue $\lambda = 0$, and (b) prove that the eigenvector $\boldsymbol{\pi}$ corresponds to a probability distribution.

(a) Since $\det(Q) = \det(Q^T)$, $\det(Q - \lambda I) = \det(Q^T - \lambda I)$, so Q and Q^T have the same eigenvalues. Since $\det(Q - \lambda I) = 0$ if and only if $\det(Q^T - \lambda I) = 0$, the left and right eigenvalues of Q are the same. However,

$$Q \times \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = 0,$$

so $\lambda = 0$ is a right eigenvalue, and hence there must be a left eigenvector with eigenvalue 0.

(b) Suppose $\boldsymbol{\pi}Q = 0$, where $\boldsymbol{\pi} = (\pi_A,\ \pi_G,\ \pi_C,\ \pi_T)$ and

$$Q = \begin{pmatrix} -(q_{12}+q_{13}+q_{14}) & q_{12} & q_{13} & q_{14} \\ q_{21} & -(q_{21}+q_{23}+q_{24}) & q_{23} & q_{24} \\ q_{31} & q_{32} & -(q_{31}+q_{32}+q_{34}) & q_{34} \\ q_{41} & q_{42} & q_{43} & -(q_{41}+q_{42}+q_{43}) \end{pmatrix}.$$

(i) Multiply the vector $\boldsymbol{\pi}$ by the first column of Q:

$$(q_{12}+q_{13}+q_{14})\pi_A = \pi_G q_{21} + \pi_C q_{31} + \pi_T q_{41}. \tag{A.1}$$

i.e. if $\pi_G,\ \pi_C$ and π_T are positive, then so is π_A.

(ii) Multiply the vector $\boldsymbol{\pi}$ by the second column of Q:

$$(q_{21} + q_{23} + q_{24})\pi_G = \pi_A q_{12} + \pi_C q_{32} + \pi_T q_{42}. \tag{A.2}$$

Now we multiply both sides of equation (A.2) by $(q_{12} + q_{13} + q_{14})$:

$$\begin{aligned}(q_{21} + q_{23} + q_{24})\pi_G(q_{12} + q_{13} + q_{14}) = q_{12}\pi_A(q_{12} + q_{13} + q_{14}) \\ + (q_{12} + q_{13} + q_{14})(\pi_C q_{32} + \pi_T q_{42}).\end{aligned} \tag{A.3}$$

Now we replace the red part in the RHS of equation (A.3) with the RHS of equation (A.1)

$$\begin{aligned}(q_{21} + q_{23} + q_{24})\pi_G(q_{12} + q_{13} + q_{14}) = q_{12}(\pi_G q_{21} + \pi_T q_{31} + \pi_T q_{41}) \\ + (q_{12} + q_{13} + q_{14})(\pi_C q_{32} + \pi_T q_{42}).\end{aligned} \tag{A.4}$$

After the $q_{21}q_{12}\pi_G$ term is cancelled out, equation (A.4) can be simplified as follows:

$$\begin{aligned}\pi_G(q_{21}(q_{13} + q_{14}) + (q_{23} + q_{24})(q_{12} + q_{13} + q_{14})) = q_{12}\pi_C q_{31} + q_{12}\pi_T q_{41} \\ + (q_{12} + q_{13} + q_{14})(\pi_C q_{32} + \pi_T q_{42}).\end{aligned} \tag{A.5}$$

Equation (A.5) can be written as

$$\pi_G A = \pi_C B + \pi_T C, \tag{A.6}$$

where A, B and C are positive quantities:

$$\begin{aligned}A &= q_{21}q_{13} + q_{21}q_{14} + q_{23}q_{12} + q_{23}q_{13} + q_{23}q_{14} + q_{24}q_{12} + q_{24}q_{13} + q_{24}q_{14}, \\ B &= q_{12}q_{31} + q_{12}q_{32} + q_{13}q_{32} + q_{14}q_{32}, \\ C &= q_{12}q_{41} + q_{12}q_{42} + q_{13}q_{42} + q_{14}q_{42},\end{aligned}$$

i.e. if π_C and π_T are positive, then so is π_G, and so is π_A.

(iii) Multiply the vector $\boldsymbol{\pi}$ by the third column of Q

$$(q_{31} + q_{32} + q_{34})\pi_C = \pi_A q_{13} + \pi_G q_{23} + \pi_T q_{43}. \tag{A.7}$$

Now we multiply both sides of equation (A.7) by $(q_{12} + q_{13} + q_{14})$:

$$\begin{aligned}(q_{31} + q_{32} + q_{34})\pi_C(q_{12} + q_{13} + q_{14}) = q_{13}\pi_A(q_{12} + q_{13} + q_{14}) \\ + (q_{12} + q_{13} + q_{14})(\pi_G q_{23} + \pi_T q_{43}).\end{aligned} \tag{A.8}$$

Now we replace the red part in equation (A.8) with the RHS of equation (A.1)

$$(q_{31}+q_{32}+q_{34})\pi_C(q_{12}+q_{13}+q_{14}) = q_{13}(\pi_G q_{21}+\pi_C q_{31}+\pi_T q_{41}) + (q_{12}+q_{13}+q_{14})(\pi_G q_{23}+\pi_T q_{43}). \tag{A.9}$$

After the $q_{31}q_{13}\pi_C$ term is cancelled out, equation (A.9) can be simplified as follows:

$$\pi_C(q_{31}(q_{12}+q_{14})+(q_{32}+q_{34})(q_{12}+q_{13}+q_{14})) = q_{13}\pi_G q_{21}+q_{13}\pi_T q_{41} + (q_{12}+q_{13}+q_{14})(\pi_G q_{23}+\pi_T q_{43}). \tag{A.10}$$

Equation (A.10) can be written as

$$\pi_C D = \pi_G E + \pi_T F, \tag{A.11}$$

where D, E and F are positive quantities:

$$\begin{aligned} D &= q_{31}q_{12}+q_{31}q_{14}+q_{32}q_{12}+q_{32}q_{13}+q_{32}q_{14}+q_{34}q_{12}+q_{34}q_{13}+q_{34}q_{14},\\ E &= q_{13}q_{21}+q_{23}q_{12}+q_{23}q_{13}+q_{23}q_{14},\\ F &= q_{13}q_{41}+q_{43}q_{12}+q_{43}q_{13}+q_{43}q_{14}. \end{aligned}$$

Now let us consider the equations (A.6) and (A.11). After multiplying the equation (A.11) by A we have

$$\pi_C AD = \pi_G AE + \pi_T AF. \tag{A.12}$$

Now we replace the red part of equation (A.12) with the RHS of equation (A.6):

$$\pi_C AD = (\pi_C B + \pi_T C)E + \pi_T AF = \pi_C BE + \pi_T(CE+AF),$$

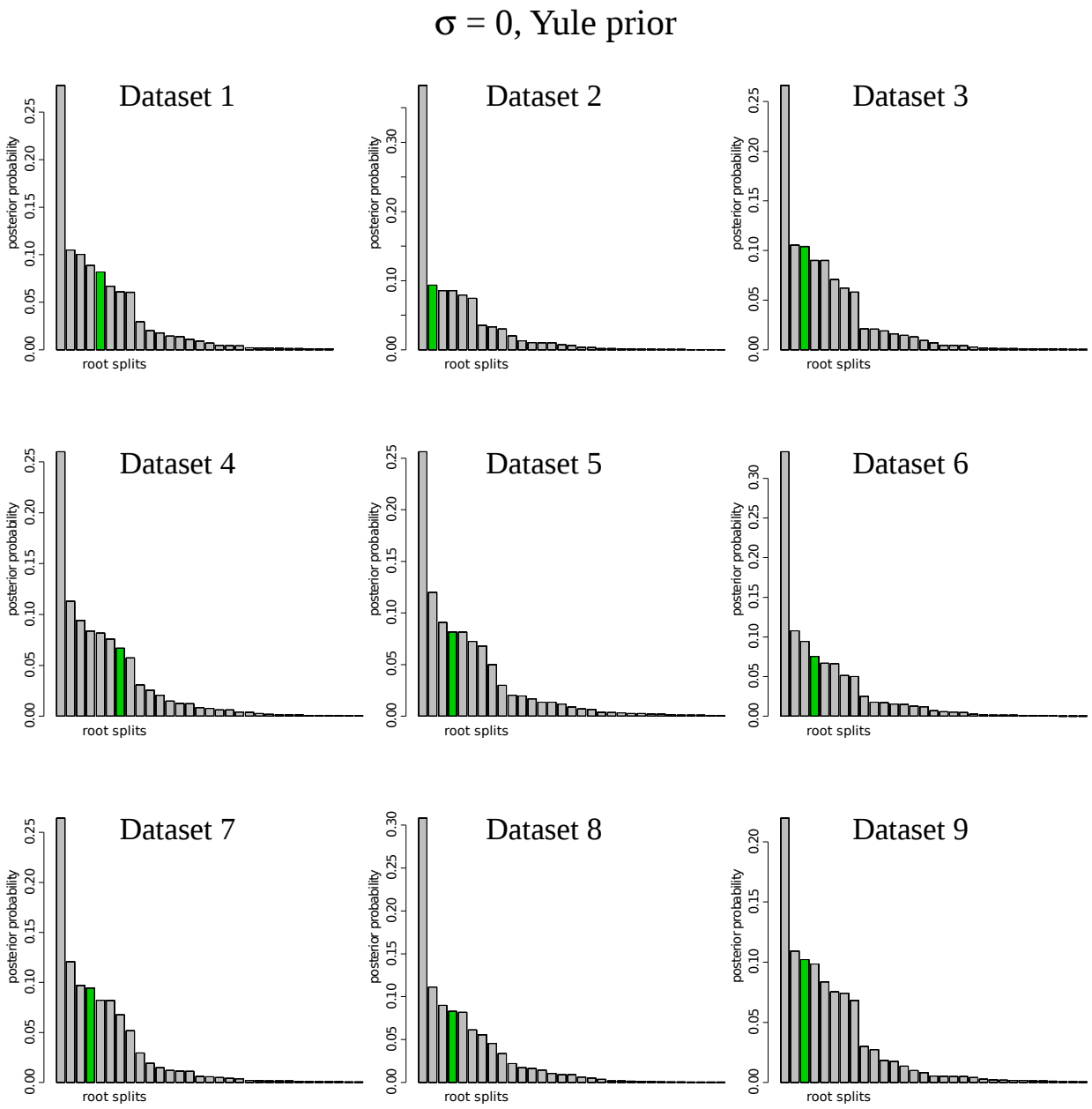
hence $\pi_C(AD-BE) = \pi_T(CE+AF)$. $AD-BE$ is positive, since $AD > BE$ (the elements with matching colours are cancelled out):

$$\begin{aligned} AD &= (q_{21}q_{13}+q_{21}q_{14}+q_{23}q_{12}+q_{23}q_{13}+q_{23}q_{14}+q_{24}q_{12}+q_{24}q_{13}+q_{24}q_{14})\\ &\quad\times(q_{31}q_{12}+q_{31}q_{14}+q_{32}q_{12}+q_{32}q_{13}+q_{32}q_{14}+q_{34}q_{12}+q_{34}q_{13}+q_{34}q_{14}).\\ BE &= (q_{13}q_{21}+q_{23}q_{12}+q_{23}q_{13}+q_{23}q_{14})\times(q_{12}q_{31}+q_{12}q_{32}+q_{13}q_{32}+q_{14}q_{32}). \end{aligned}$$

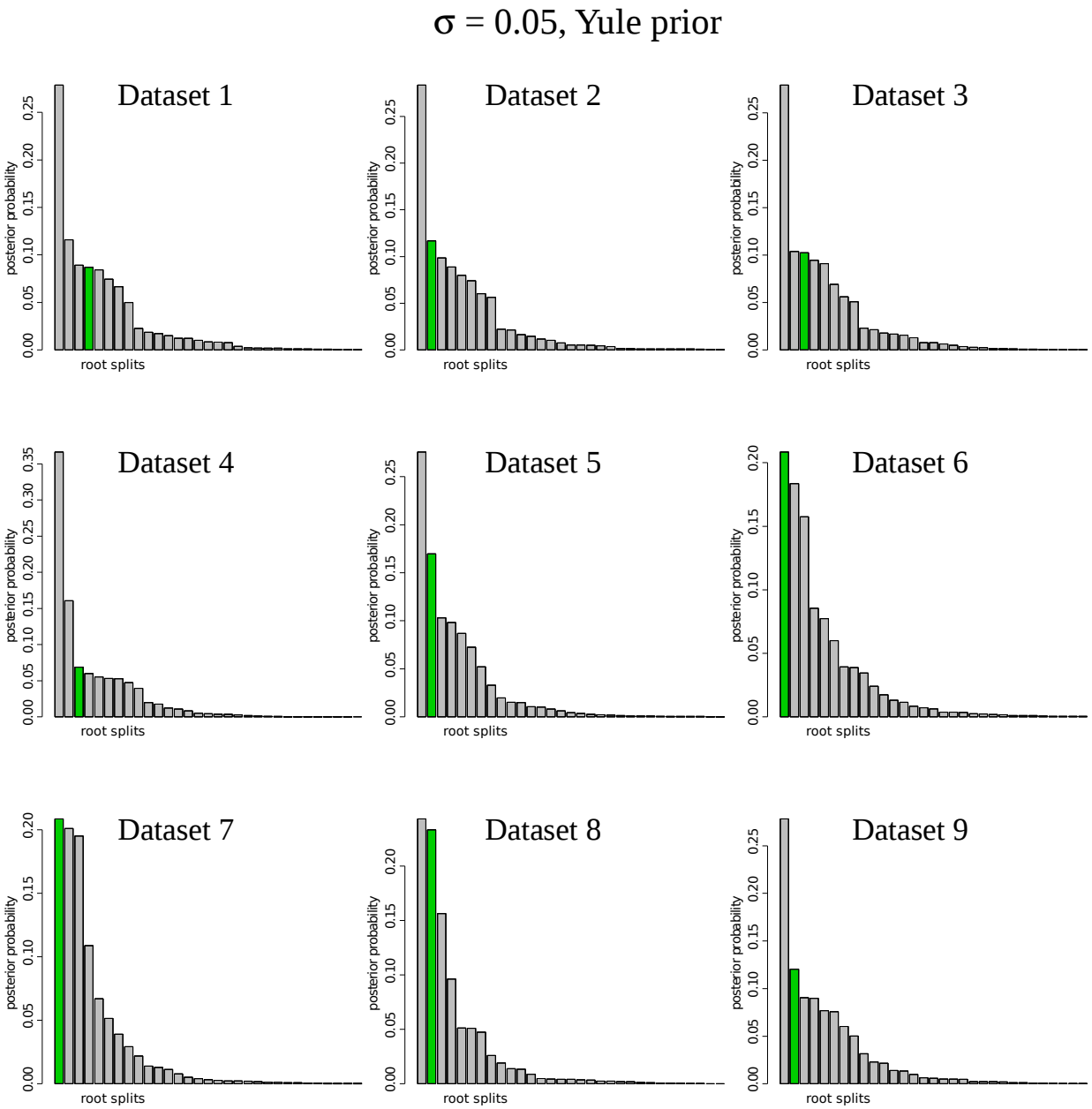
i.e. if π_T is positive, then so is π_C, and so are π_G and π_A. Thus we have proved the existence of the left eigenvector $\boldsymbol{\pi}$ whose elements are either all positive or all negative, so $\boldsymbol{\pi}$ corresponds to a probability distribution.

Appendix B

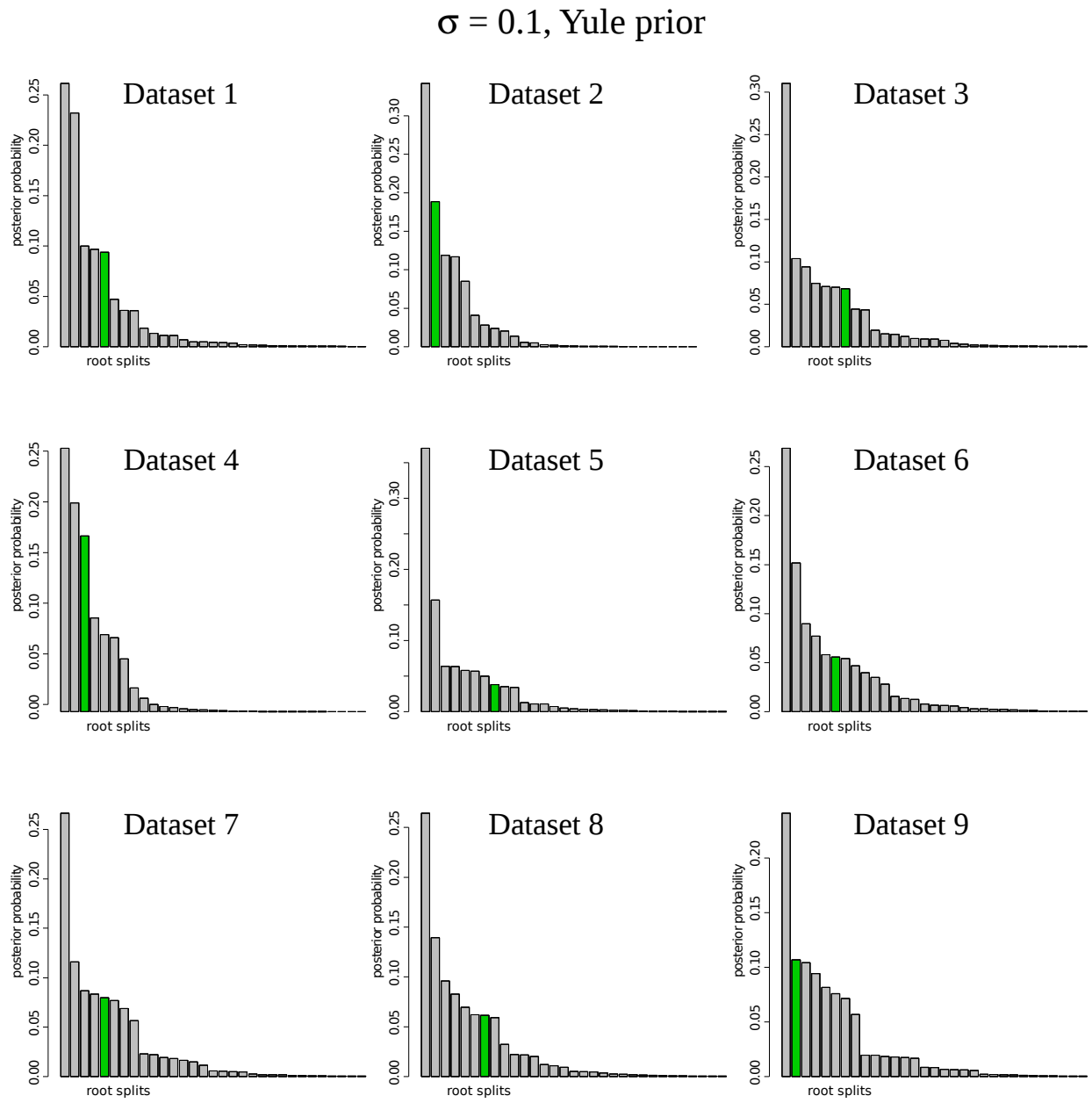
This appendix summarises the results of the first block of simulations for the NR model, analysed with the Yule prior. Figure B.1 shows the posterior distribution of the root splits for $\sigma = 0,\ 0.05,\ 0.1,\ 0.2,\ 0.3$. Different bars on the plots represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). On each plot the green bar represents the true root split. Figure B.2 shows the posterior distribution of the unrooted topologies for $\sigma = 0,\ 0.05,\ 0.1,\ 0.2,\ 0.3$. Different bars on the plots represent different unrooted topologies (ordered by posterior probabilities). On each plot the green bar represents the true unrooted topology. Each subfigure contains an analysis of nine alignments simulated with a particular value of σ.



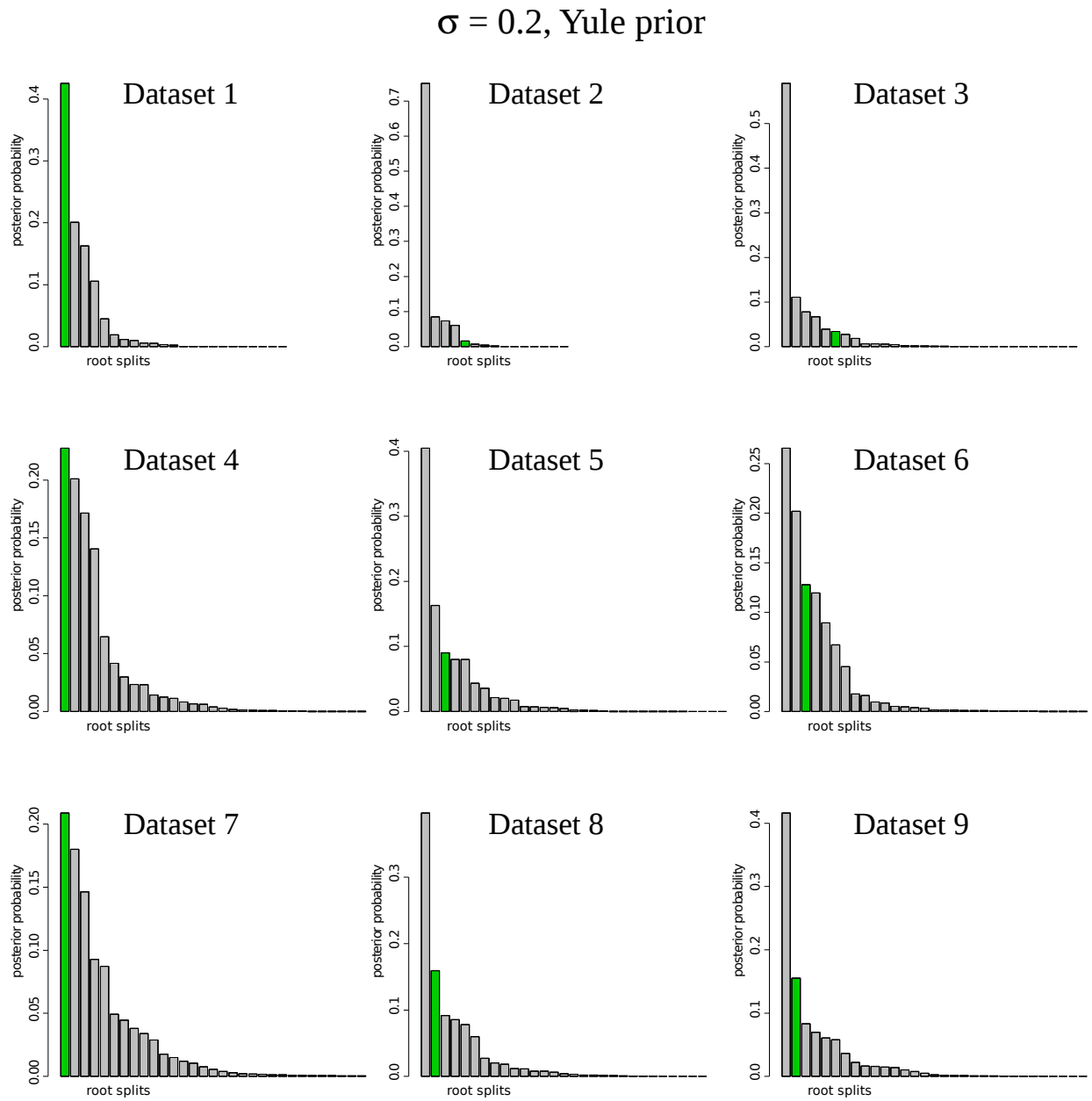
(a) Yule prior, $\sigma = 0$.



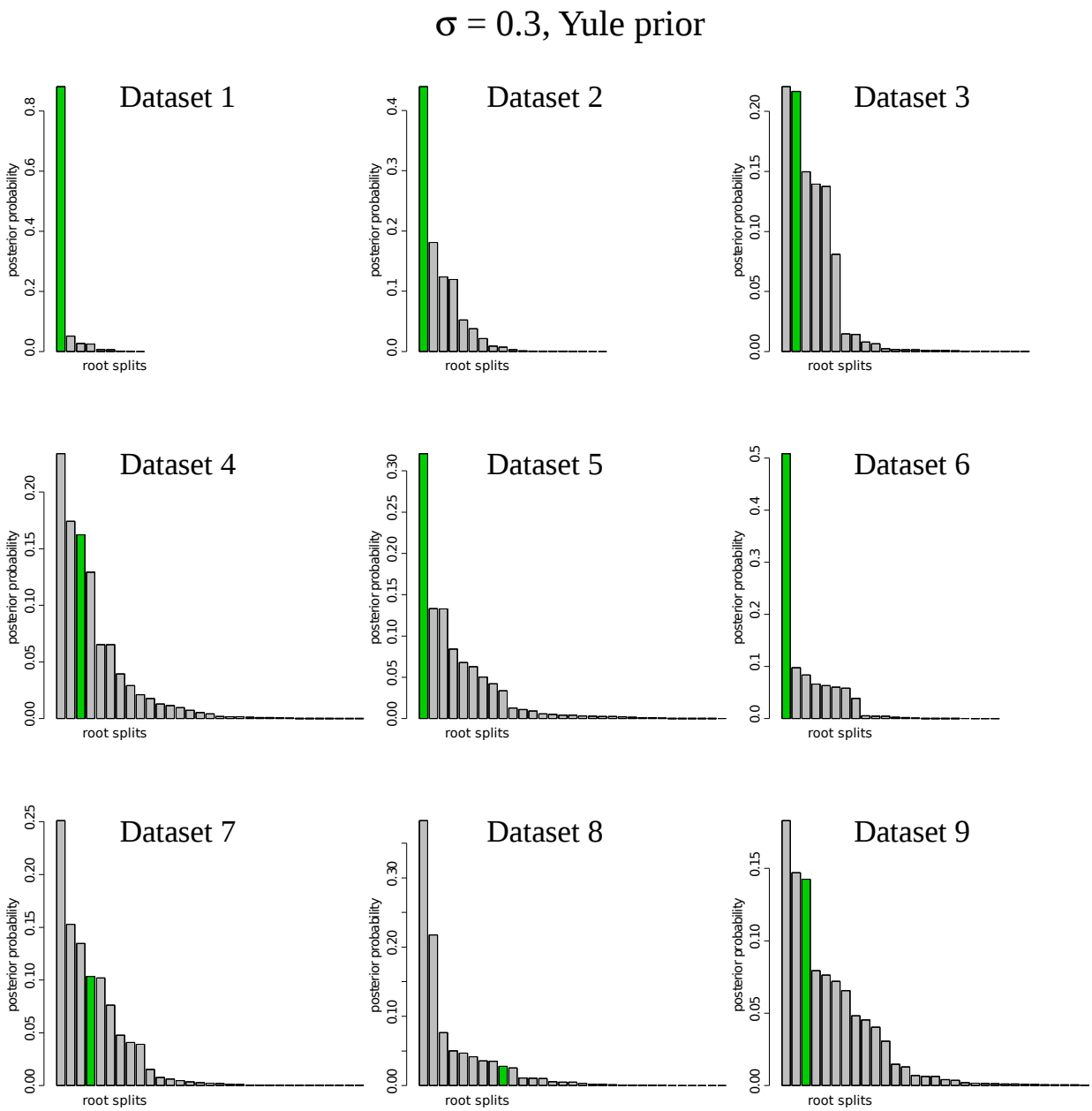
(b) Yule prior, $\sigma = 0.05$.



(c) Yule prior, $\sigma = 0.1$.

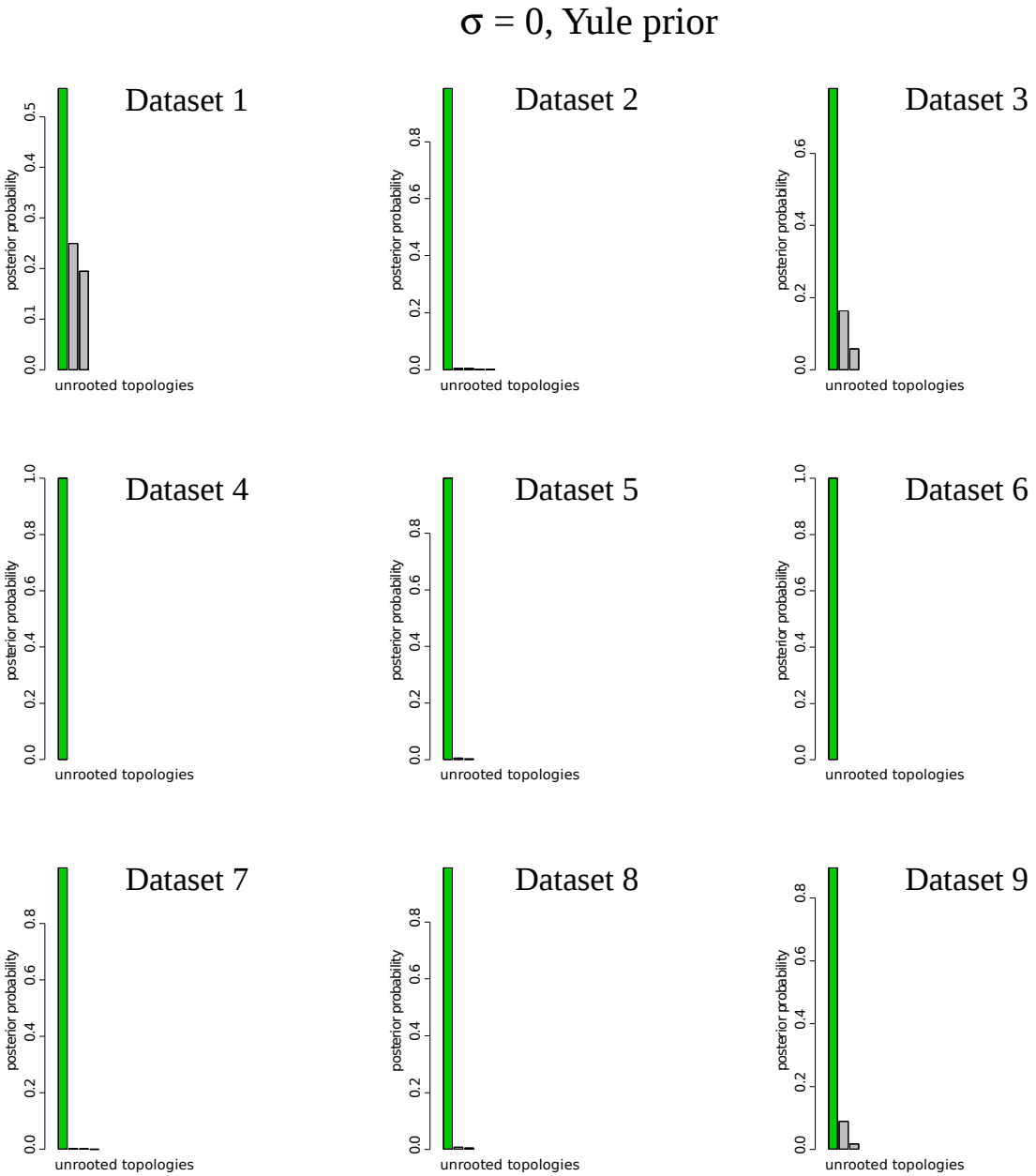


(d) Yule prior, $\sigma = 0.2$.

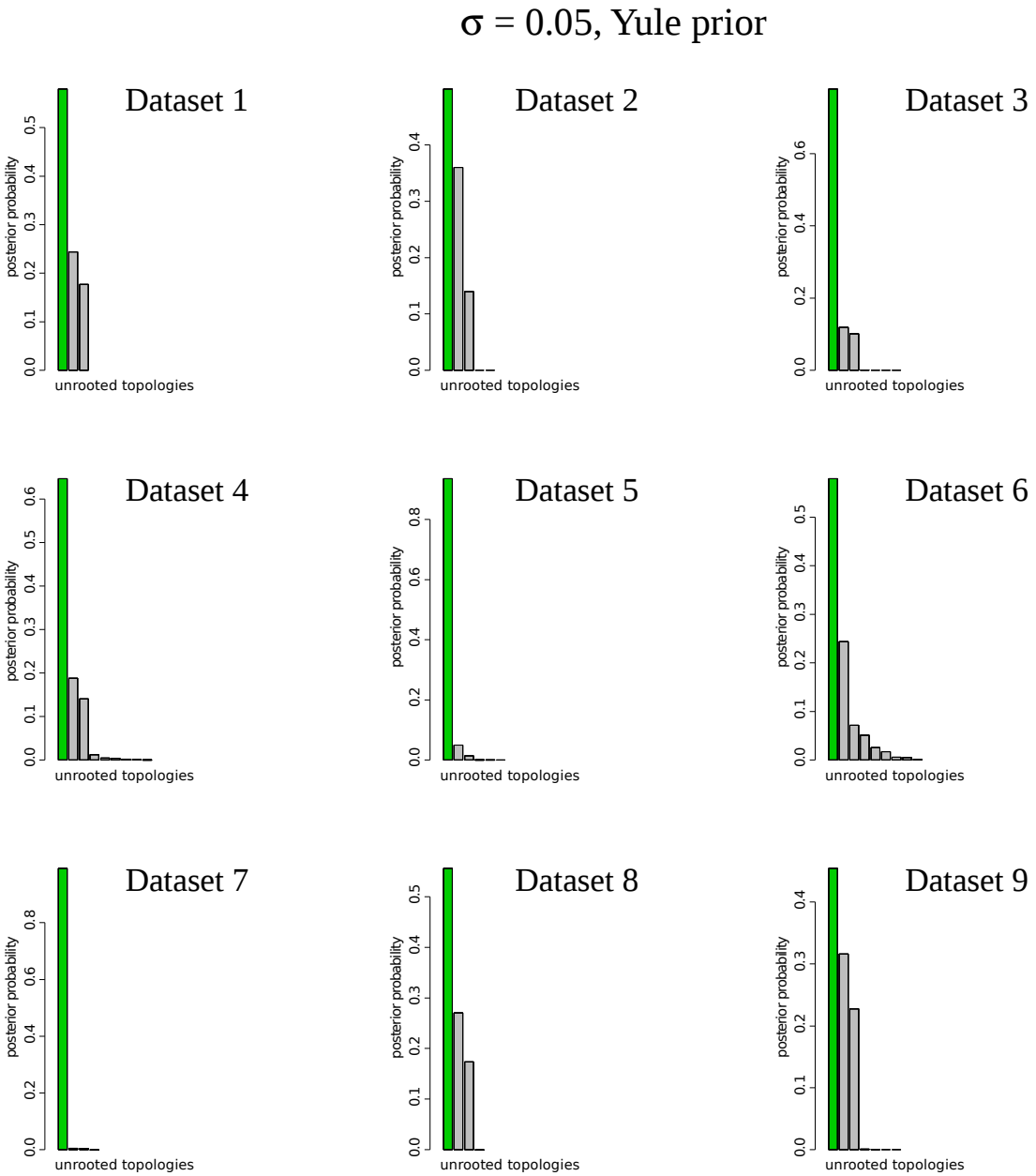


(e) Yule prior, $\sigma = 0.3$.

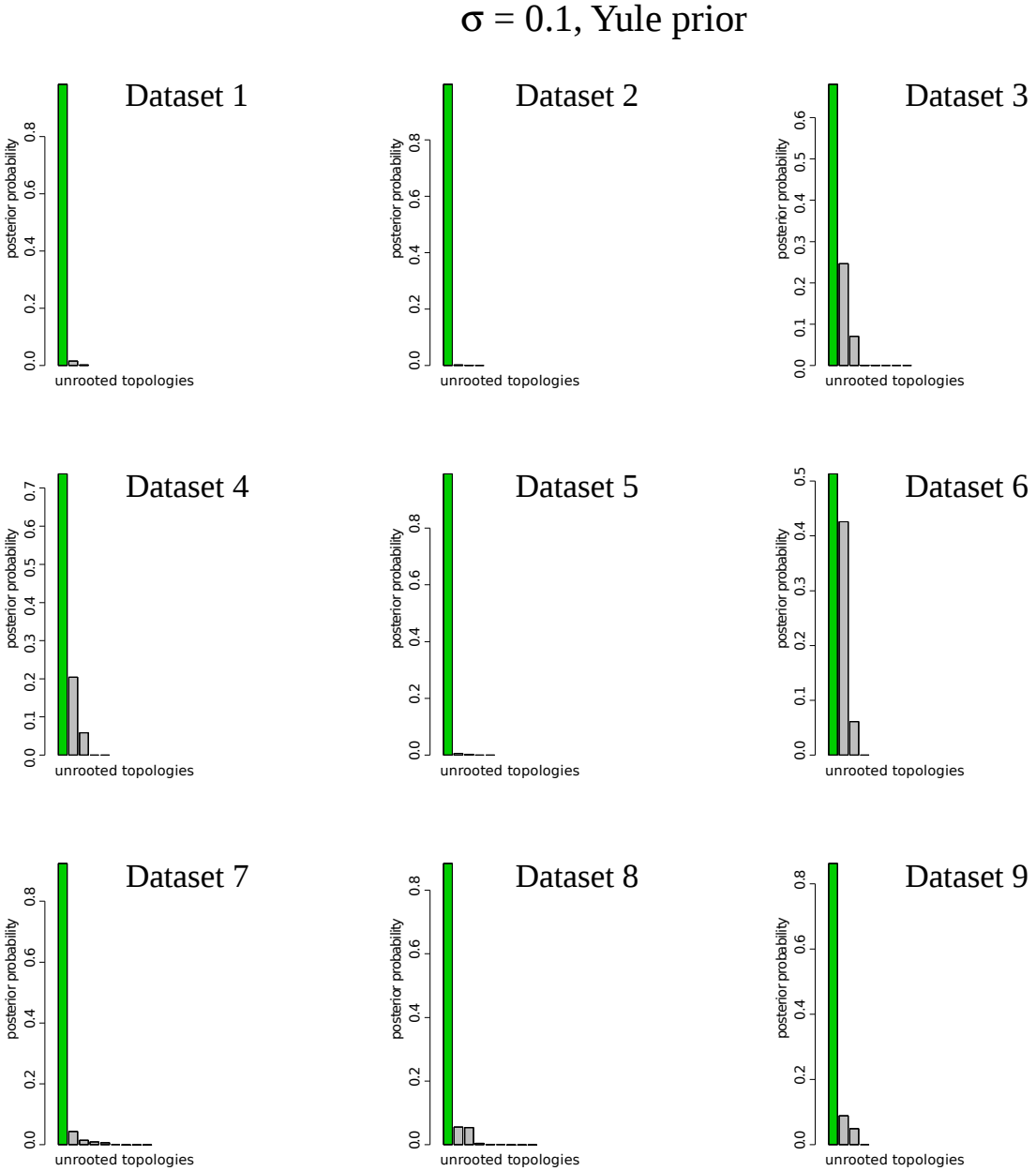
Figure B.1: Posterior distribution of the root splits for different values of σ and the Yule prior.



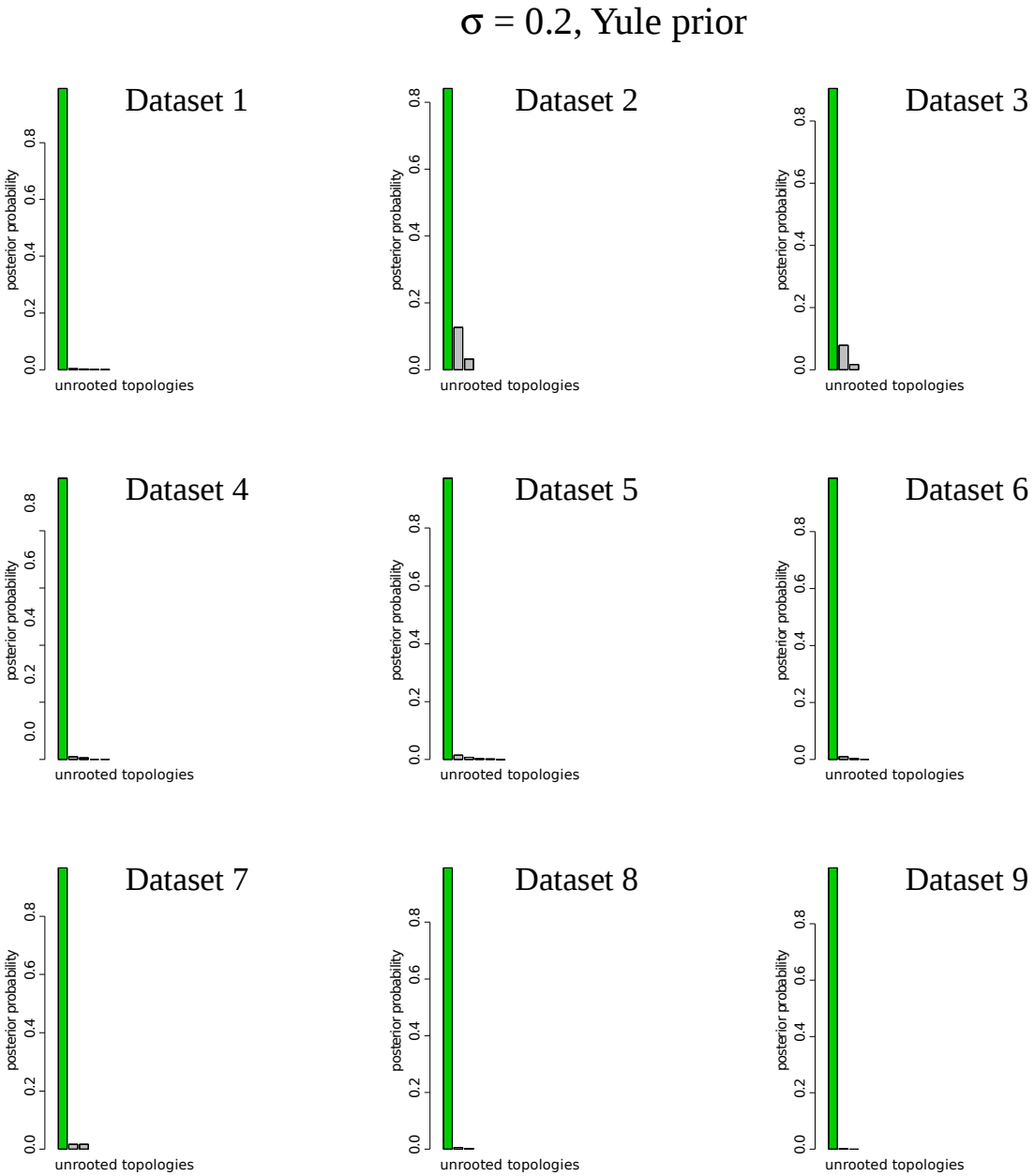
(a) Yule prior, $\sigma = 0$.



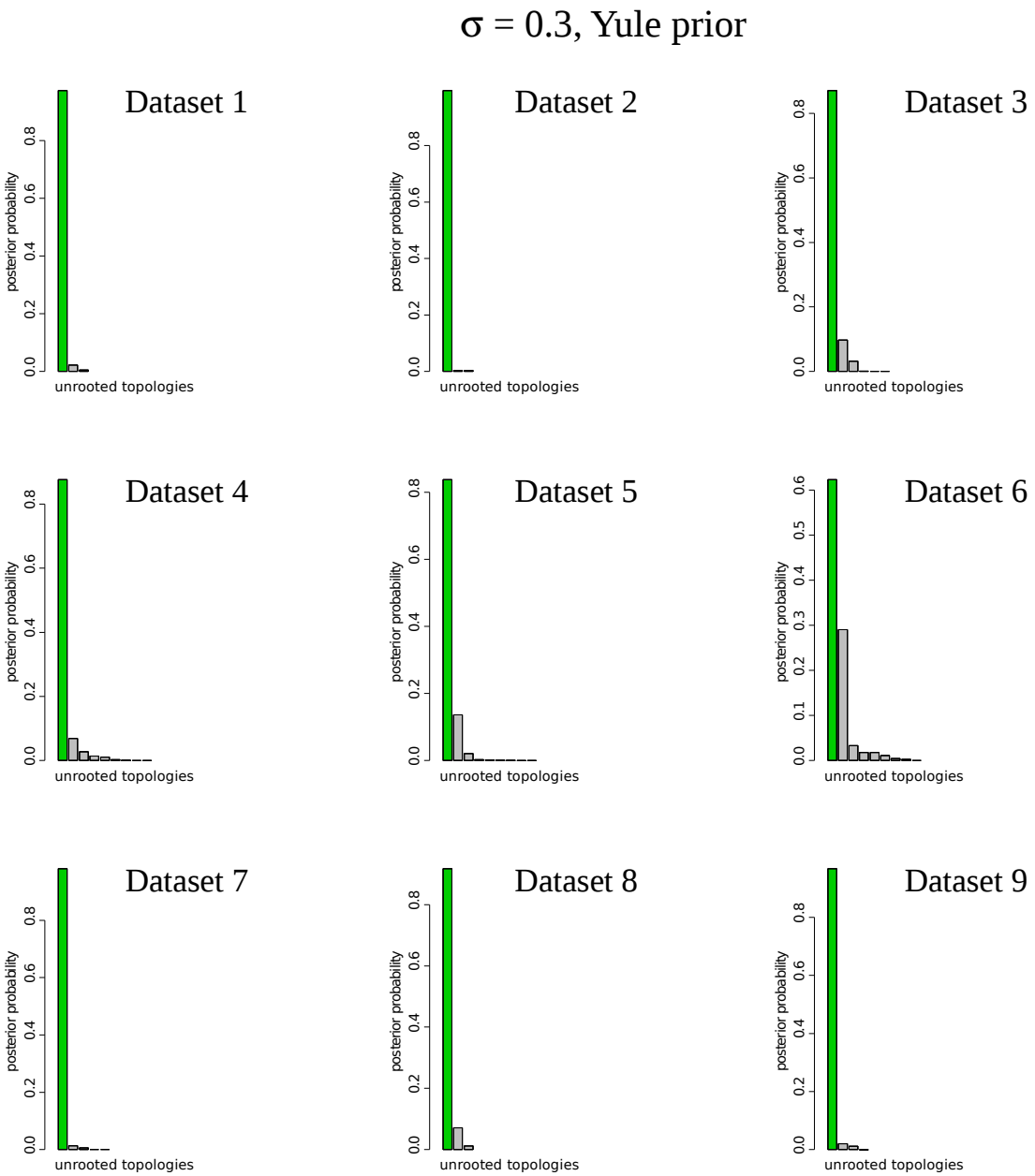
(b) Yule prior, $\sigma = 0.05$.



(c) Yule prior, $\sigma = 0.1$.



(d) Yule prior, $\sigma = 0.2$.

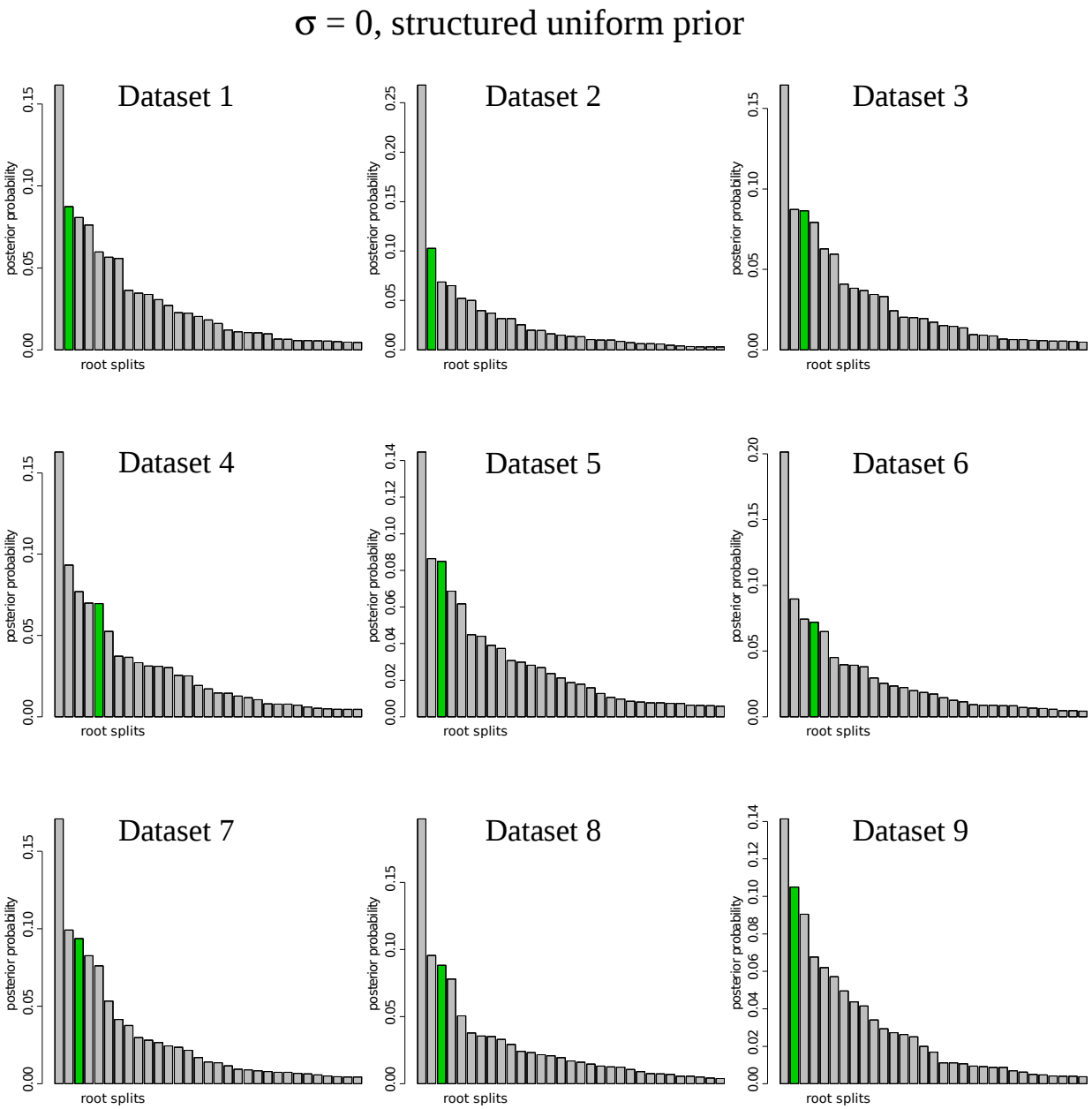


(e) Yule prior, $\sigma = 0.3$.

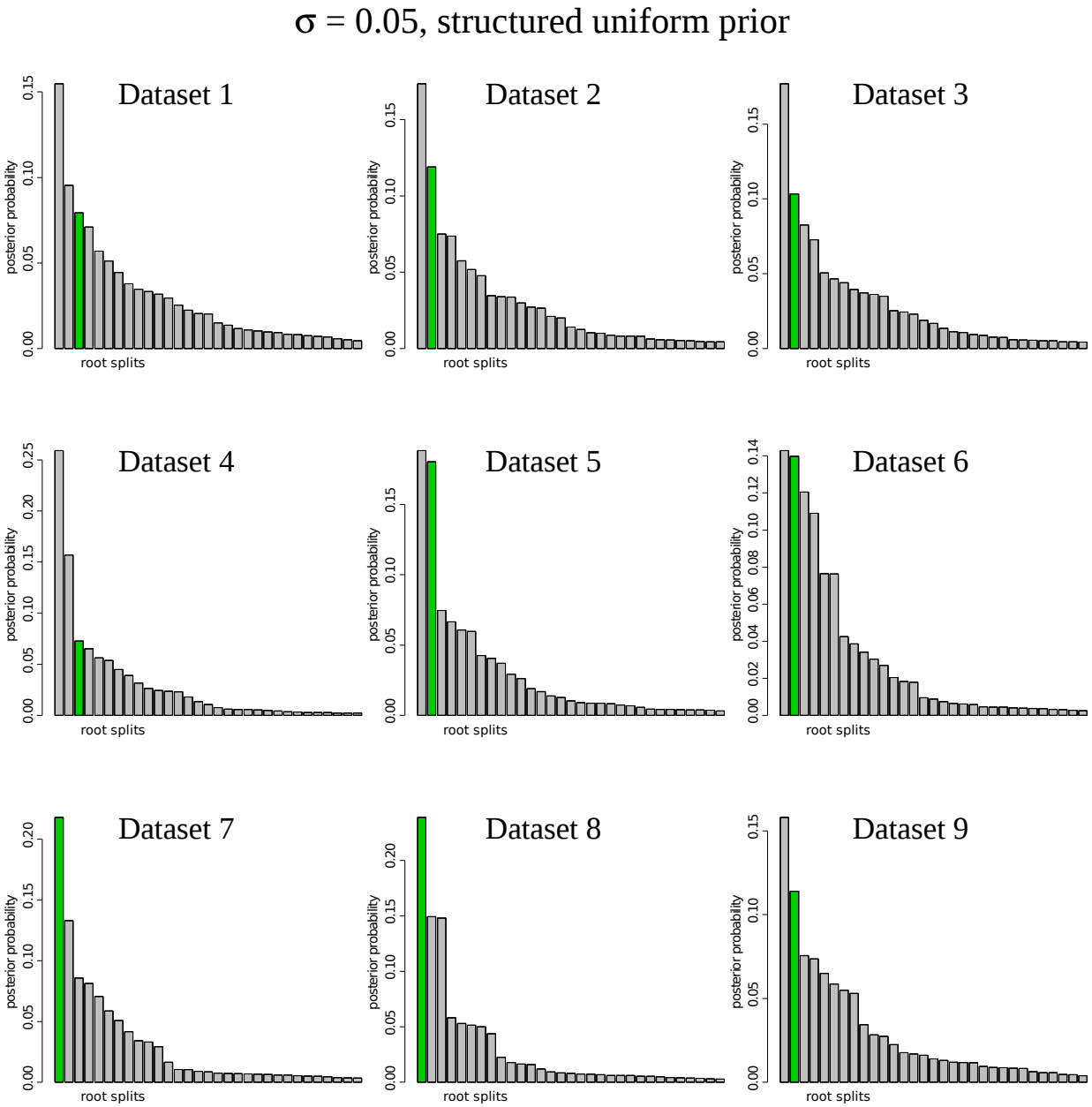
Figure B.2: Posterior distribution of the unrooted topologies for different values of σ and the Yule prior.

Appendix C

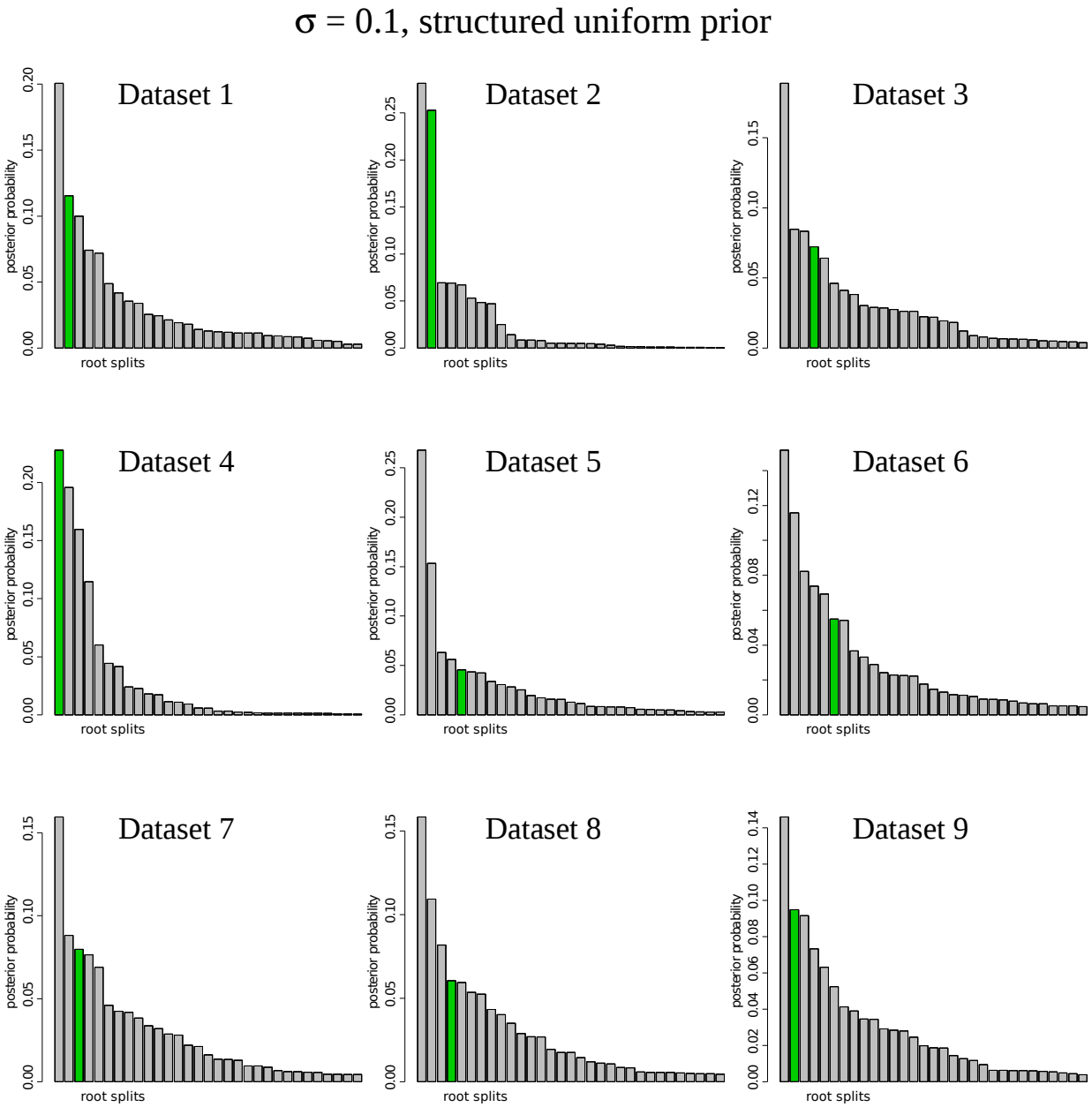
This appendix summarises the results of the first block of simulations for the NR model, analysed with the structured uniform prior. Figure C.1 shows the posterior distribution of the root splits for $\sigma = 0,\ 0.05,\ 0.1,\ 0.2,\ 0.3$. Different bars on the plots represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). On each plot the green bar represents the true root split. Figure C.2 shows the posterior distribution of the unrooted topologies for $\sigma = 0,\ 0.05,\ 0.1,\ 0.2,\ 0.3$. Different bars on the plots represent different unrooted topologies (ordered by posterior probabilities). On each plot the green bar represents the true unrooted topology. Each subfigure contains an analysis of nine alignments simulated with a particular value of σ.



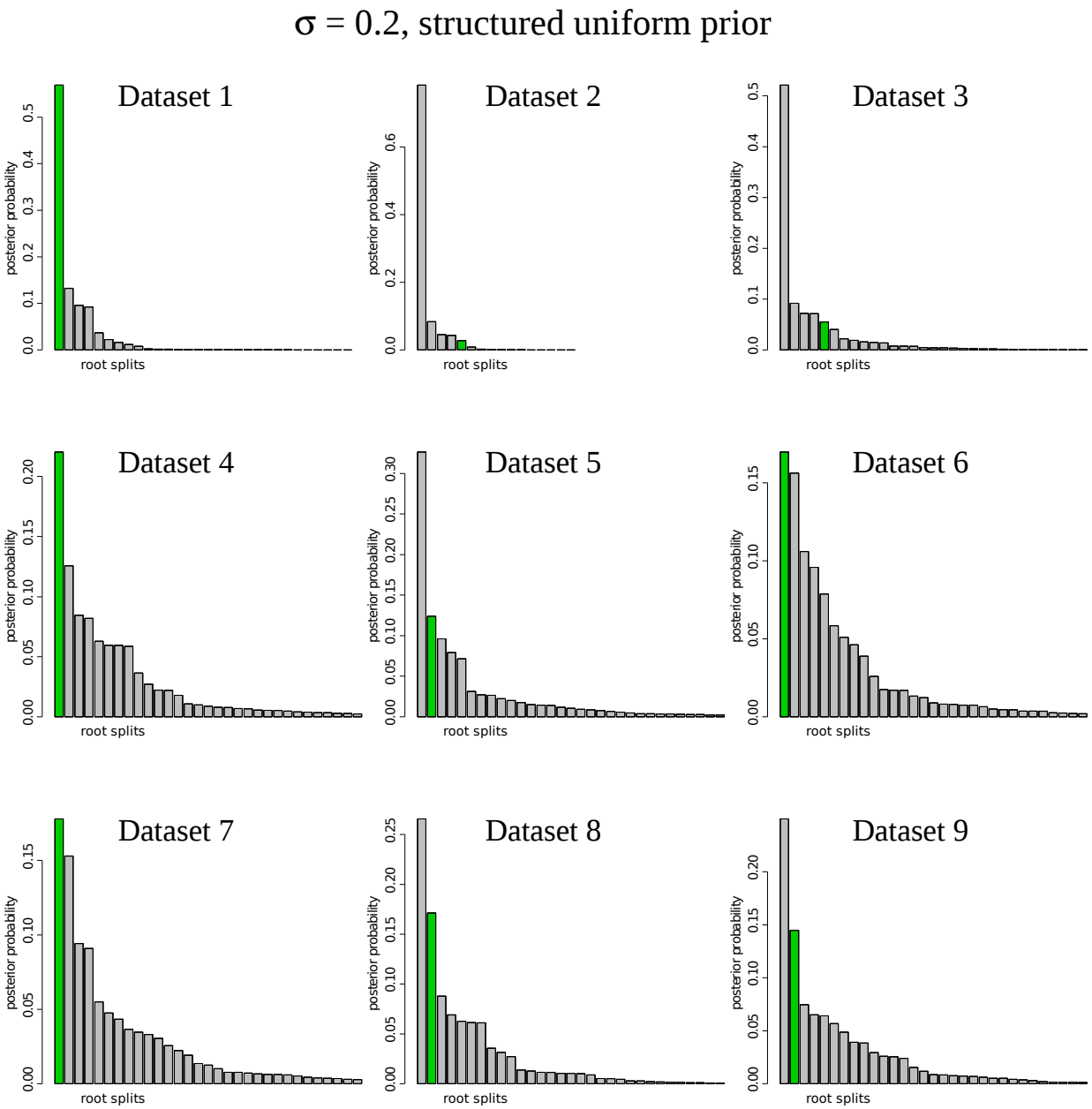
(a) Structured uniform prior, $\sigma = 0$.



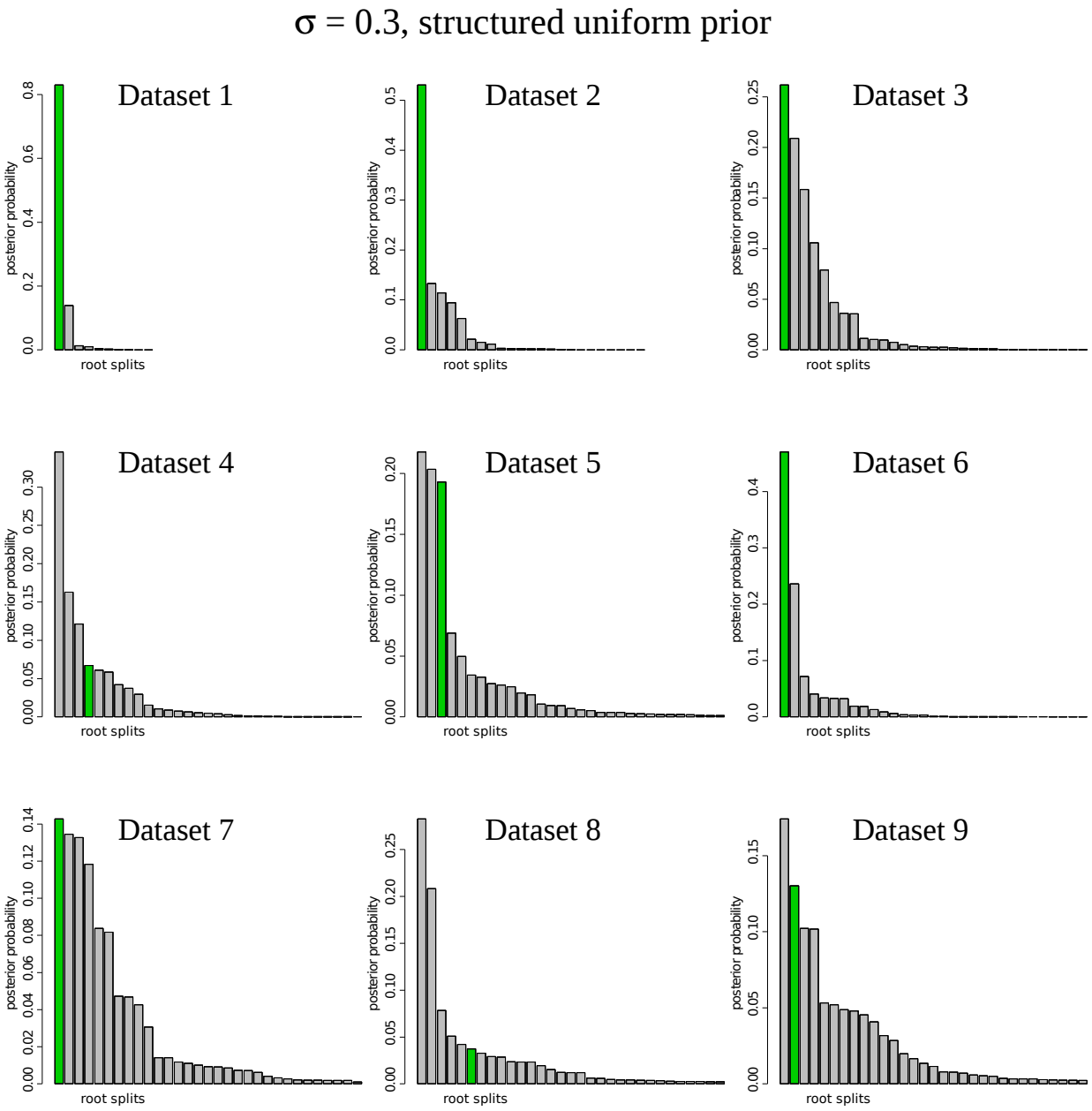
(b) Structured uniform prior, $\sigma = 0.05$.



(c) Structured uniform prior, $\sigma = 0.1$.

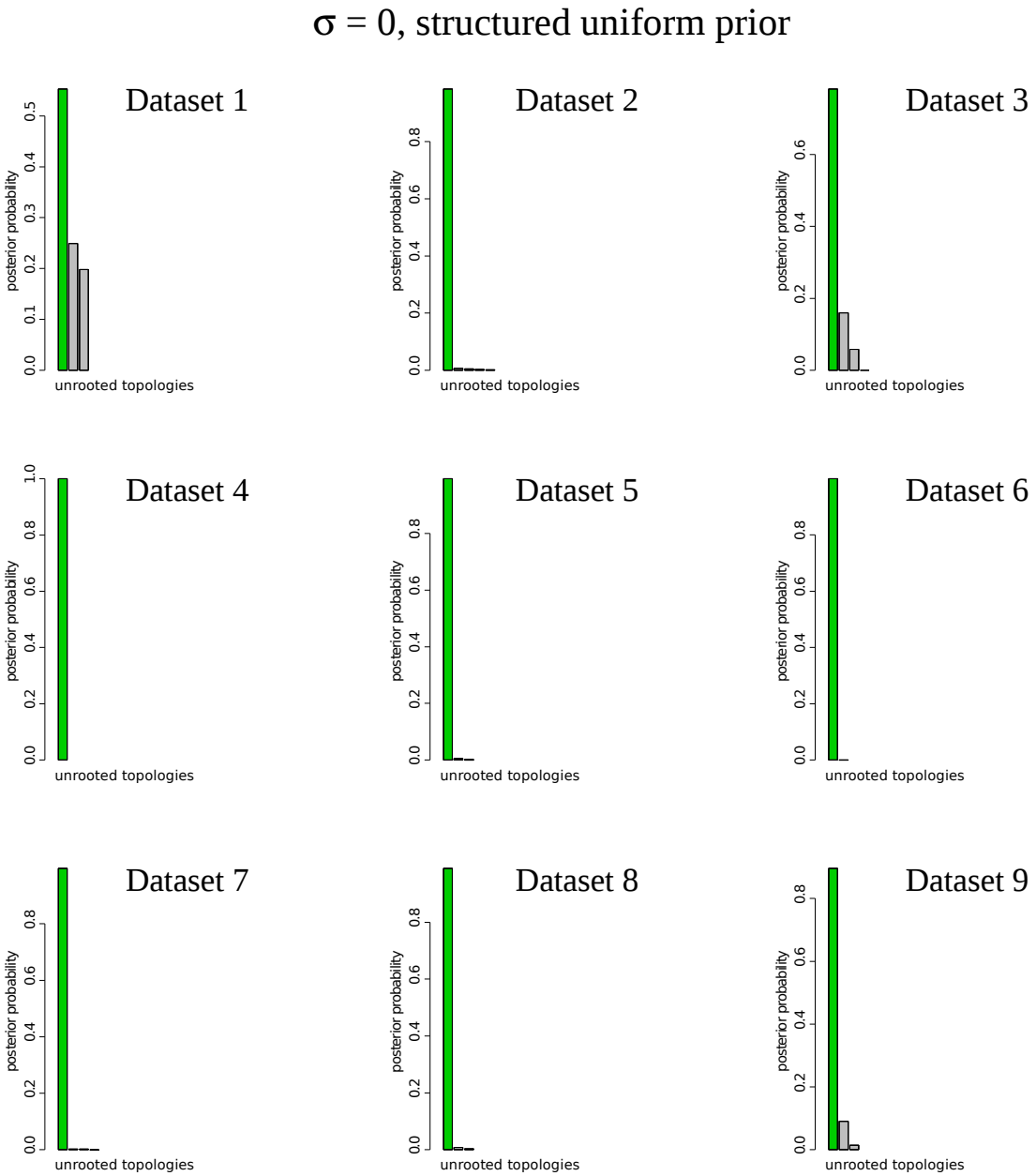


(d) Structured uniform prior, $\sigma = 0.2$.

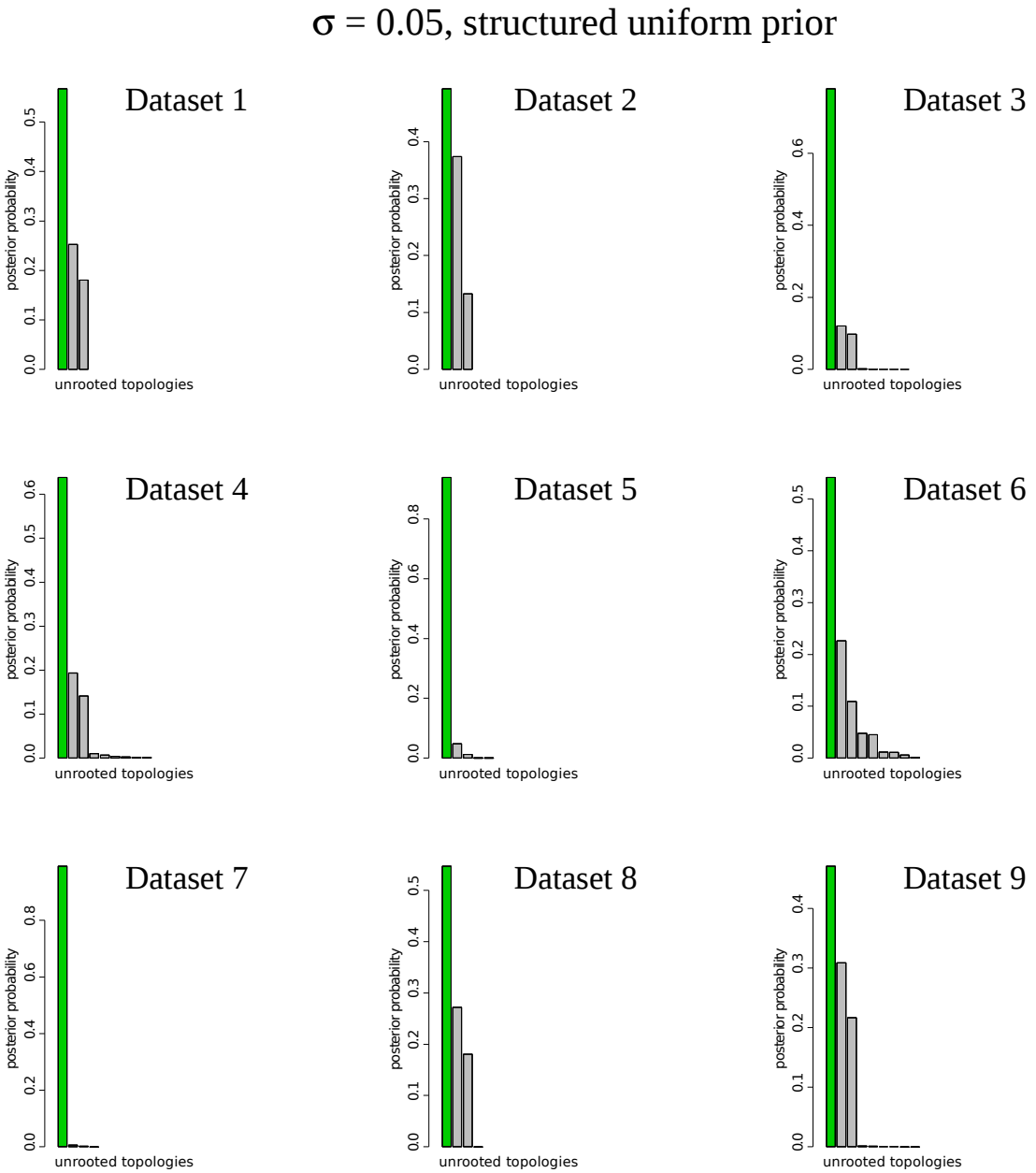


(e) Structured uniform prior, $\sigma = 0.3$.

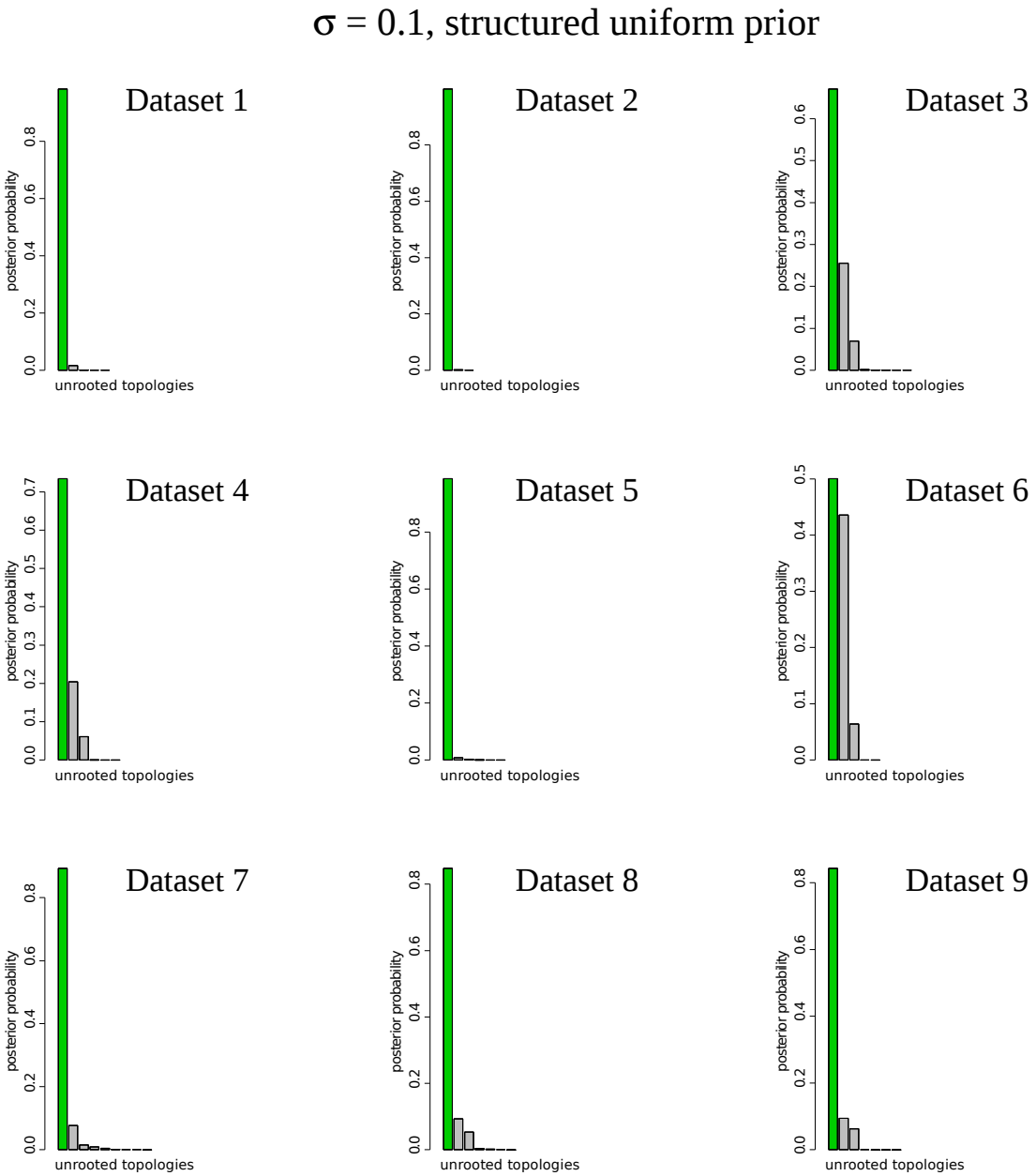
Figure C.1: Posterior distribution of the root splits for different values of σ and structured uniform prior.



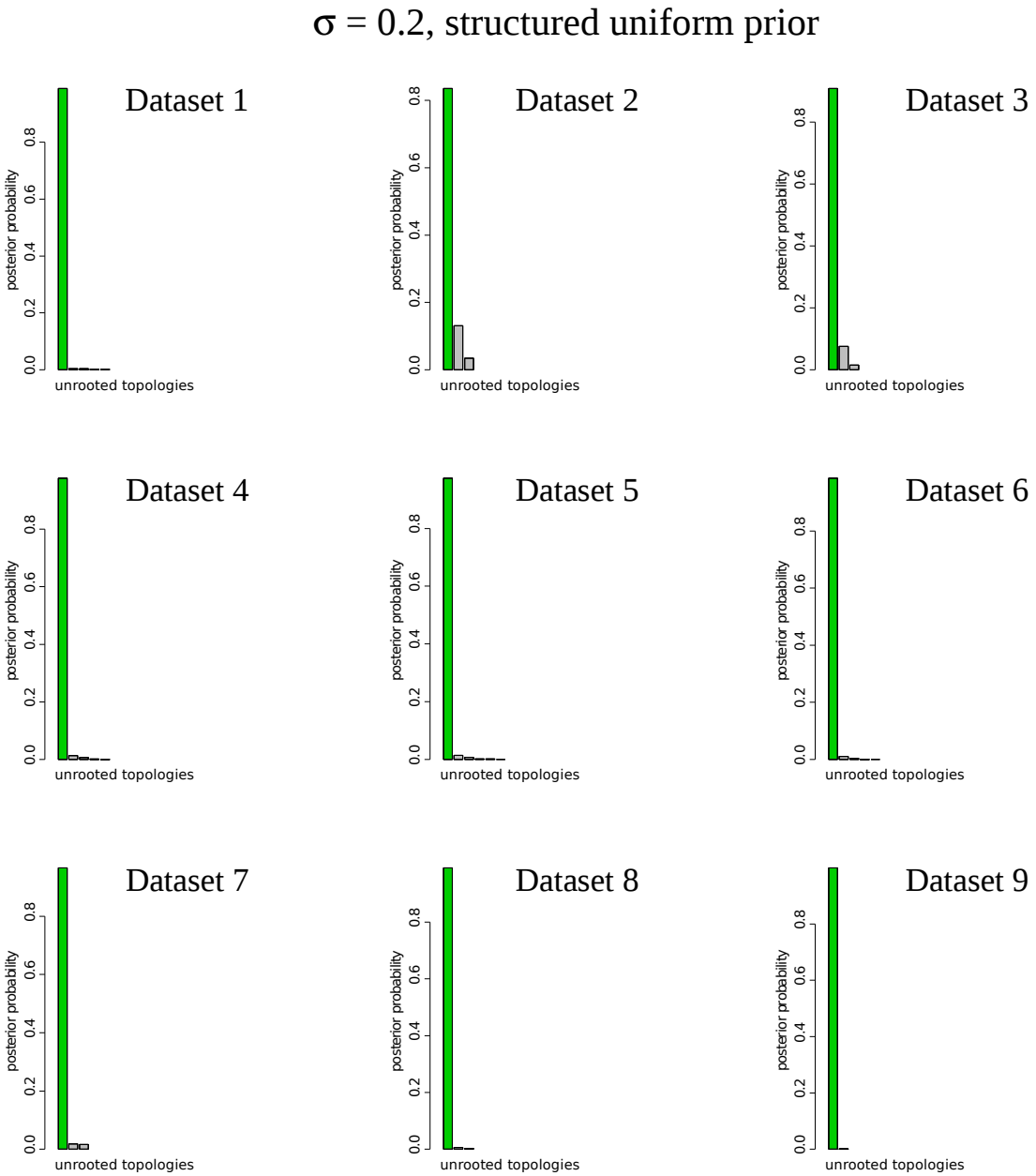
(a) Structured uniform prior, $\sigma = 0$.



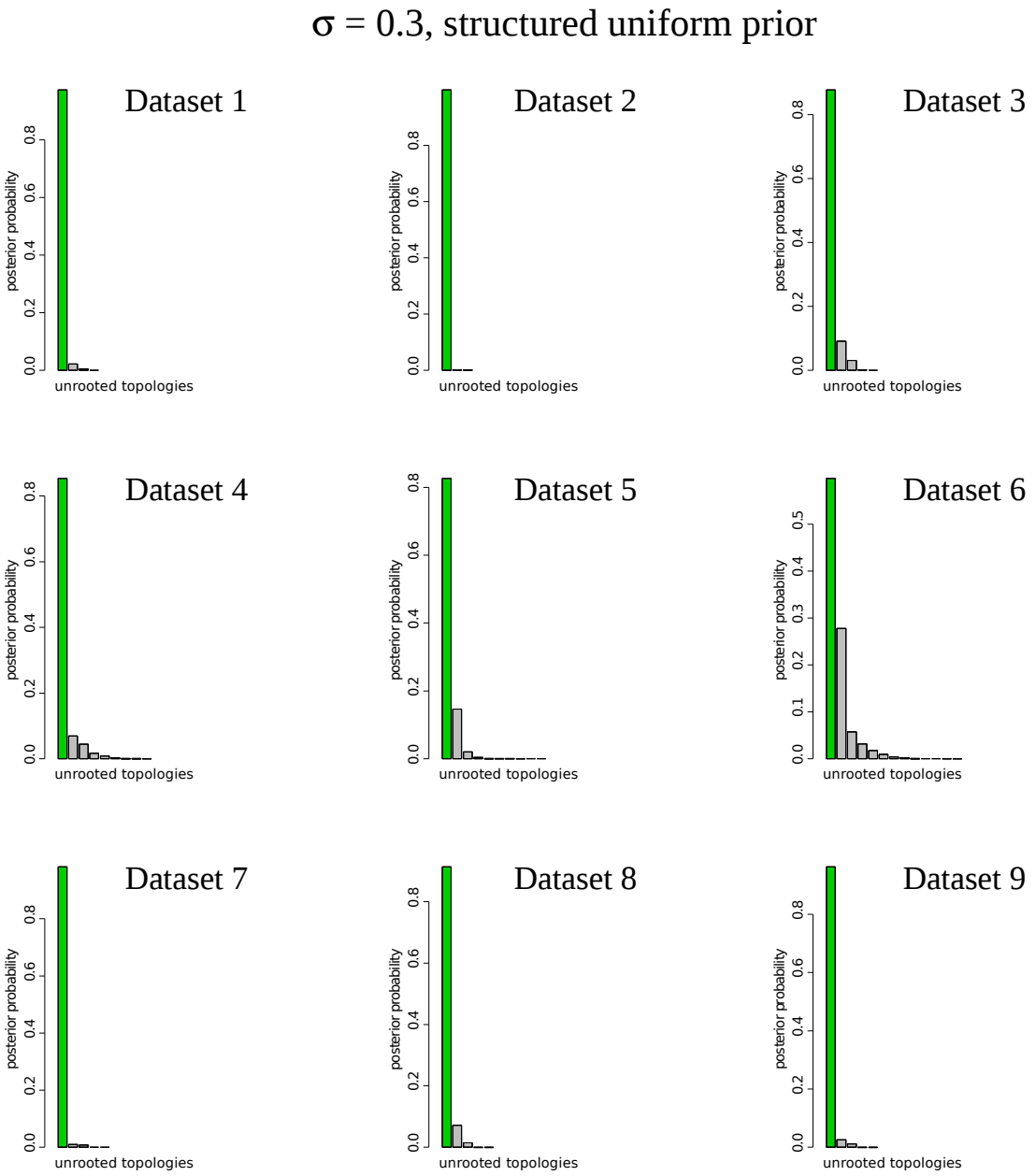
(b) Structured uniform prior, $\sigma = 0.05$.



(c) Structured uniform prior, $\sigma = 0.1$.



(d) Structured uniform prior, $\sigma = 0.2$.

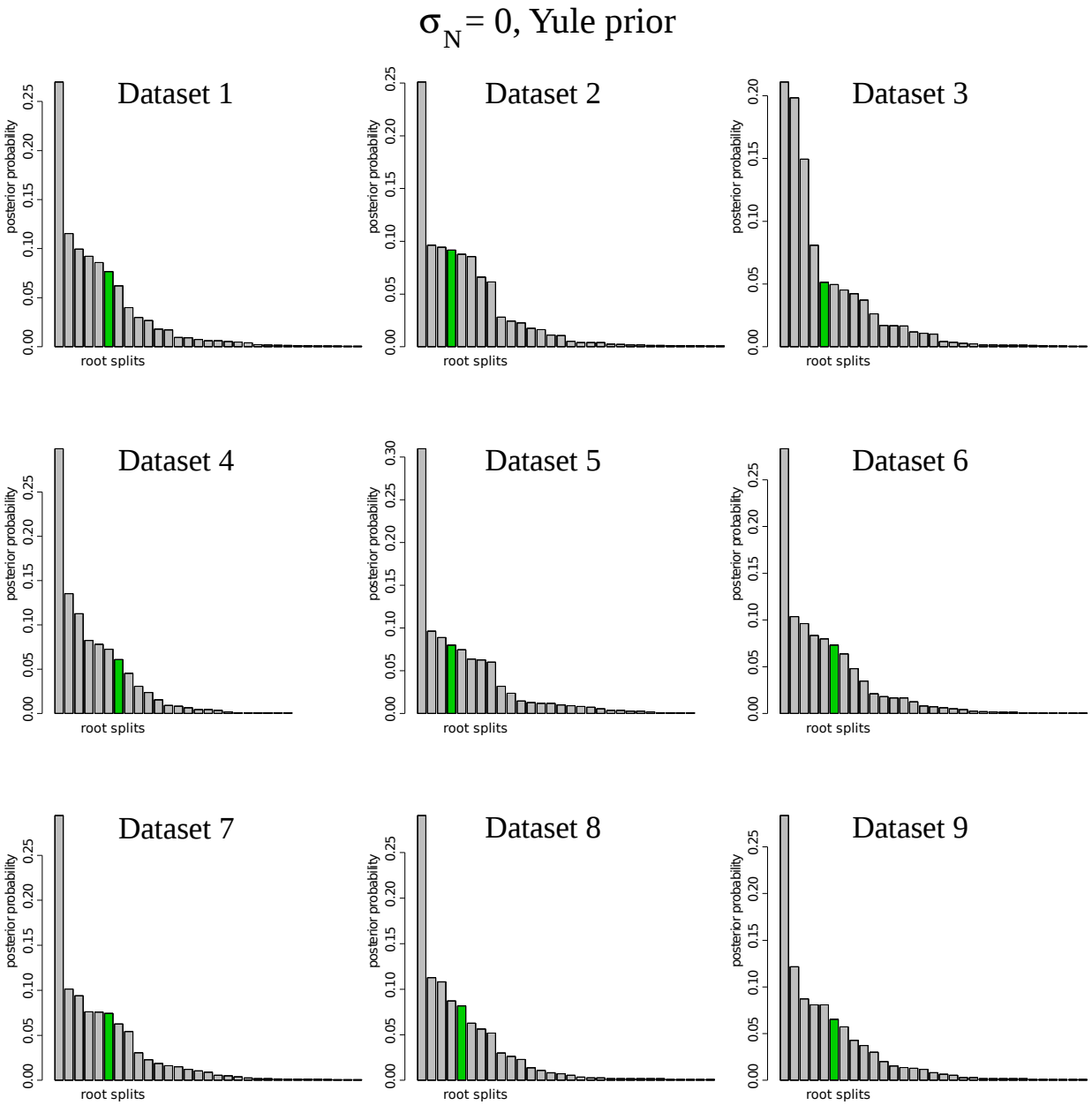


(e) Structured uniform prior, $\sigma = 0.3$.

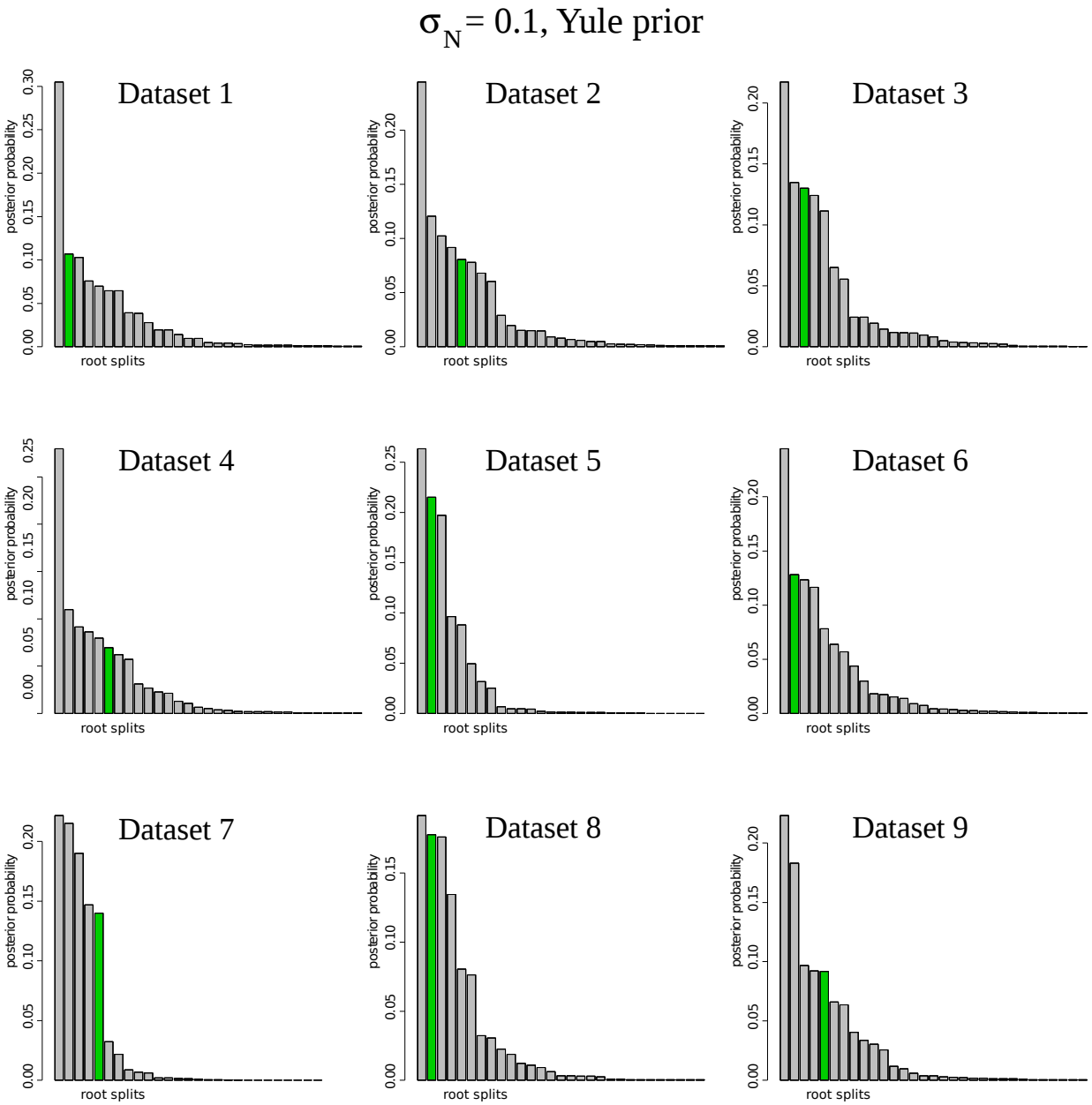
Figure C.2: Posterior distribution of the unrooted topologies for different values of σ and structured uniform prior.

Appendix D

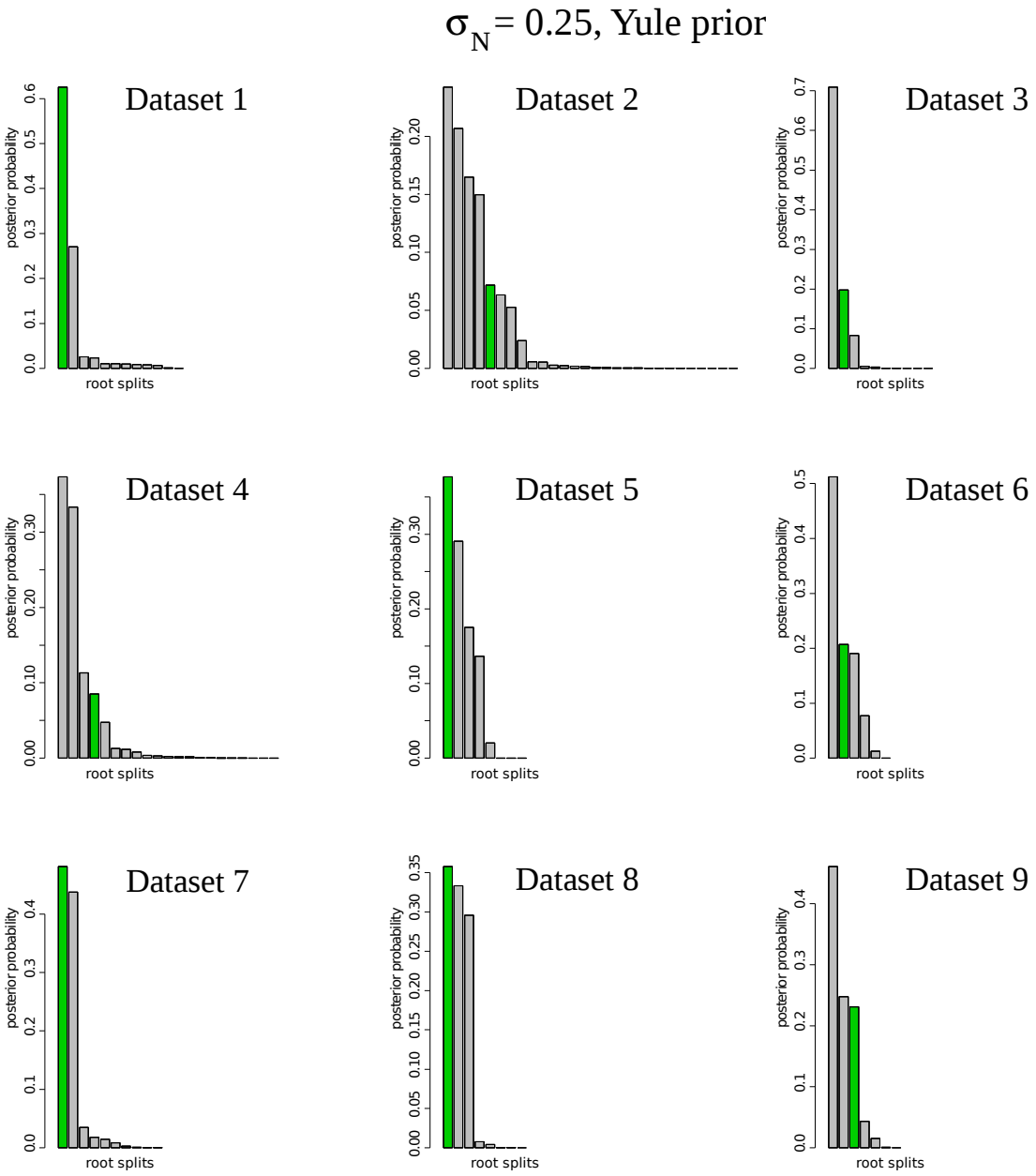
This appendix summarises the results of the first block of simulations for the NR2 model, analysed with the Yule prior. Figure D.1 shows the posterior distribution of the root splits for $\sigma_N = 0,\ 0.1,\ 0.25,\ 0.5,\ 1$. Different bars on the plots represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). On each plot the green bar represents the true root split. Figure D.2 shows the posterior distribution of the unrooted topologies for $\sigma_N = 0,\ 0.1,\ 0.25,\ 0.5,\ 1$. Different bars on the plots represent different unrooted topologies (ordered by posterior probabilities). On each plot the green bar represents the true unrooted topology. Each subfigure contains an analysis of nine alignments simulated with a particular value of σ_N.



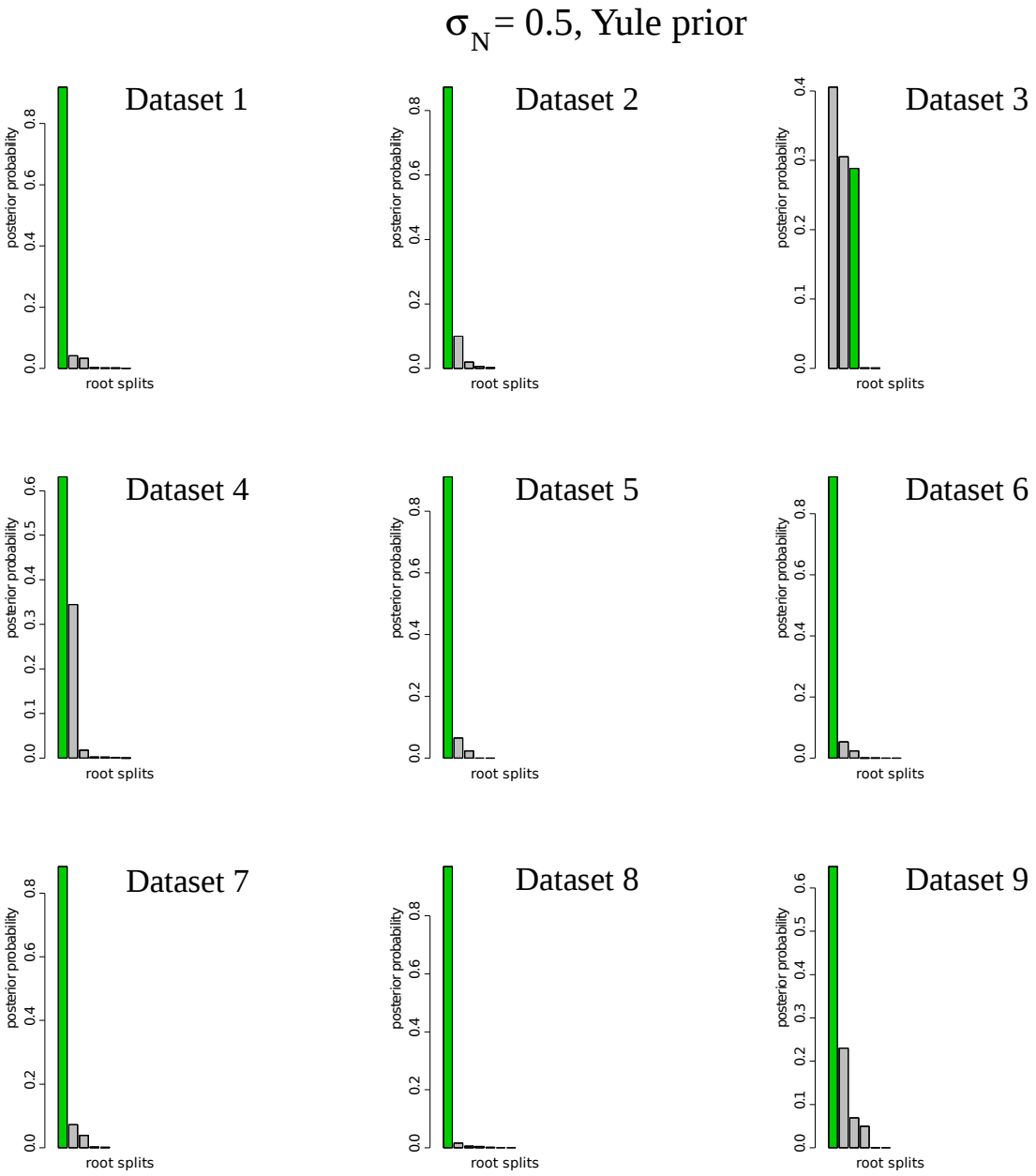
(a) Yule prior, $\sigma_N = 0$.



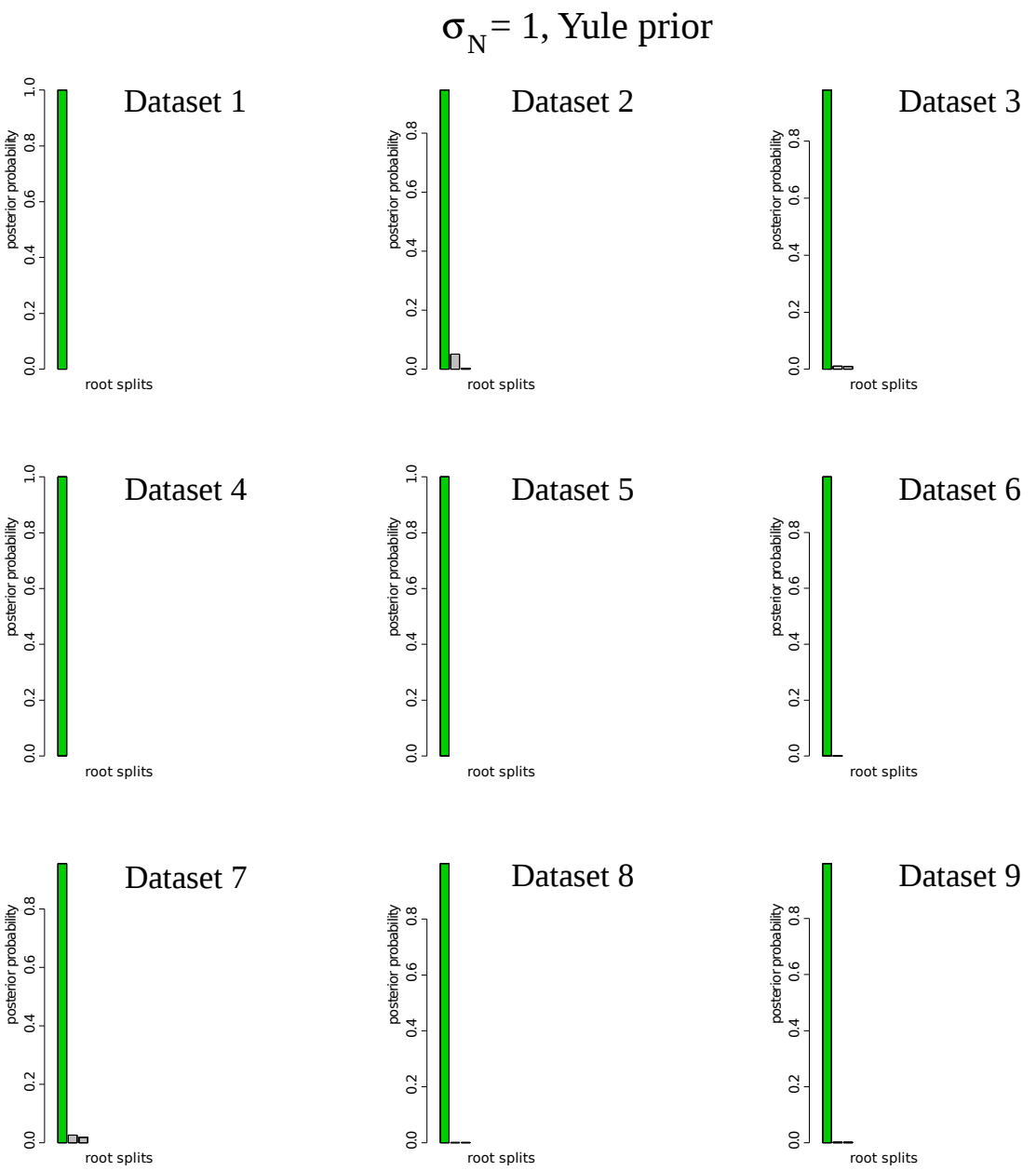
(b) Yule prior, $\sigma_N = 0.1$.



(c) Yule prior, $\sigma_N = 0.25$.

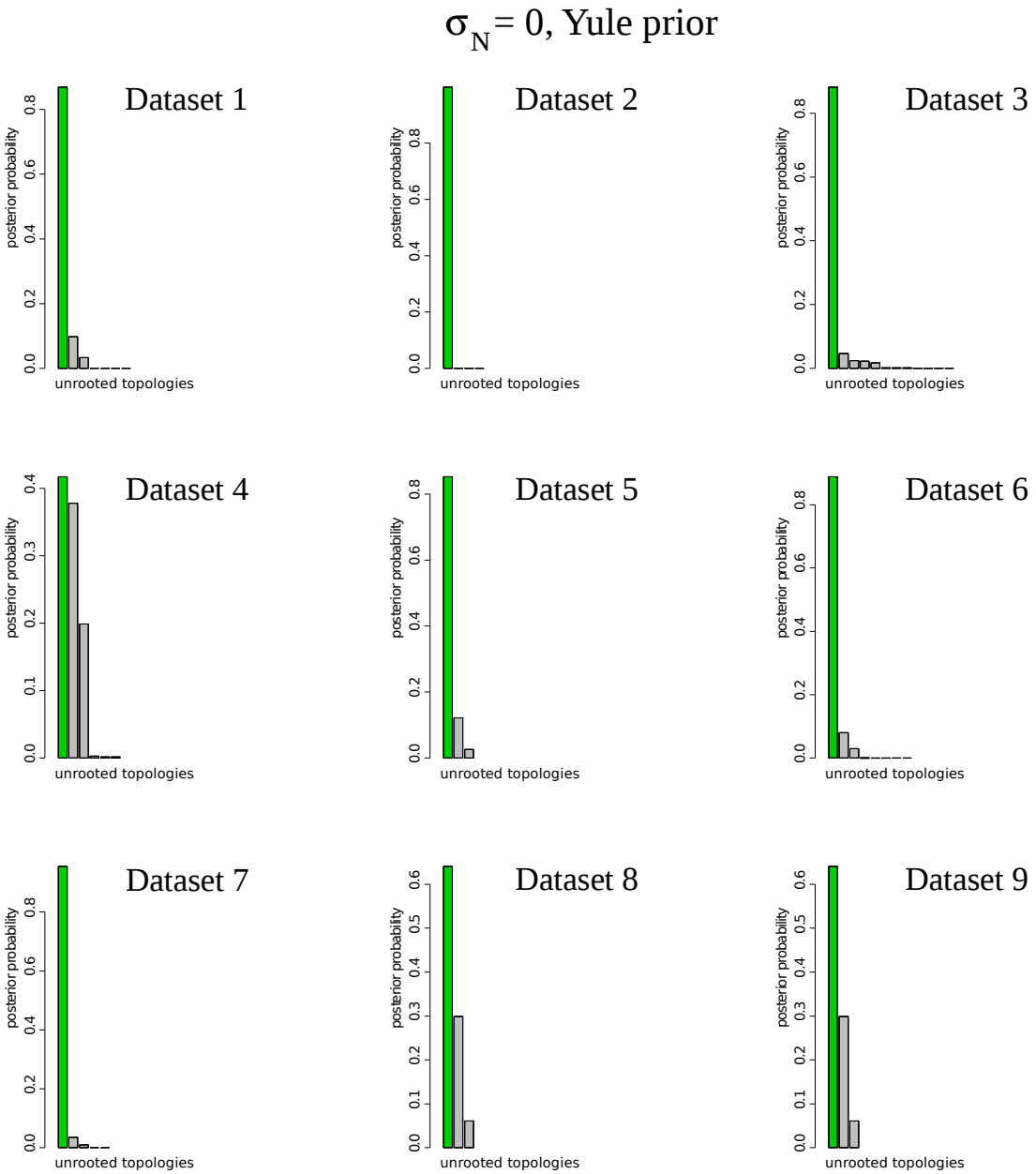


(d) Yule prior, $\sigma_N = 0.5$.

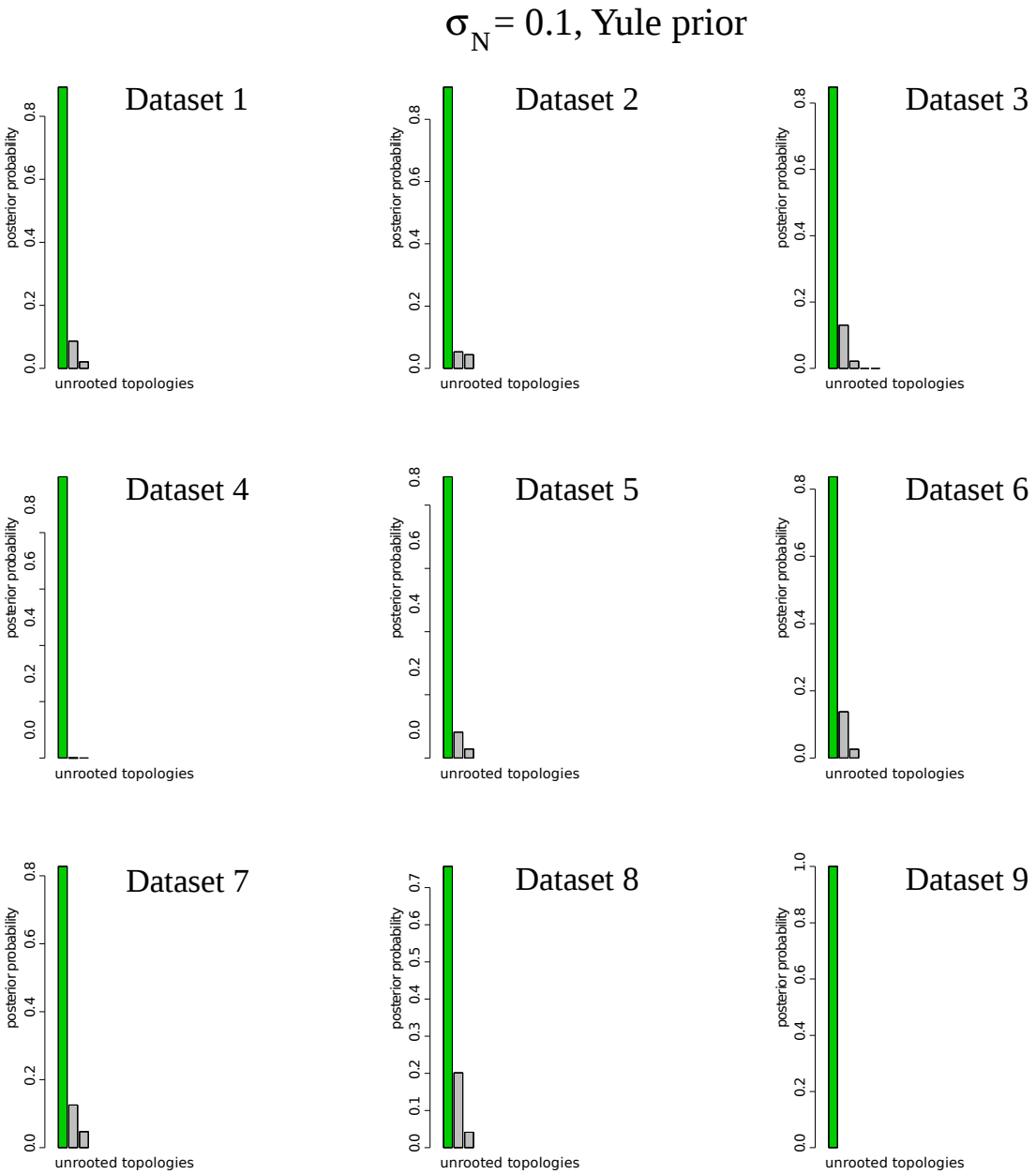


(e) Yule prior, $\sigma_N = 1$.

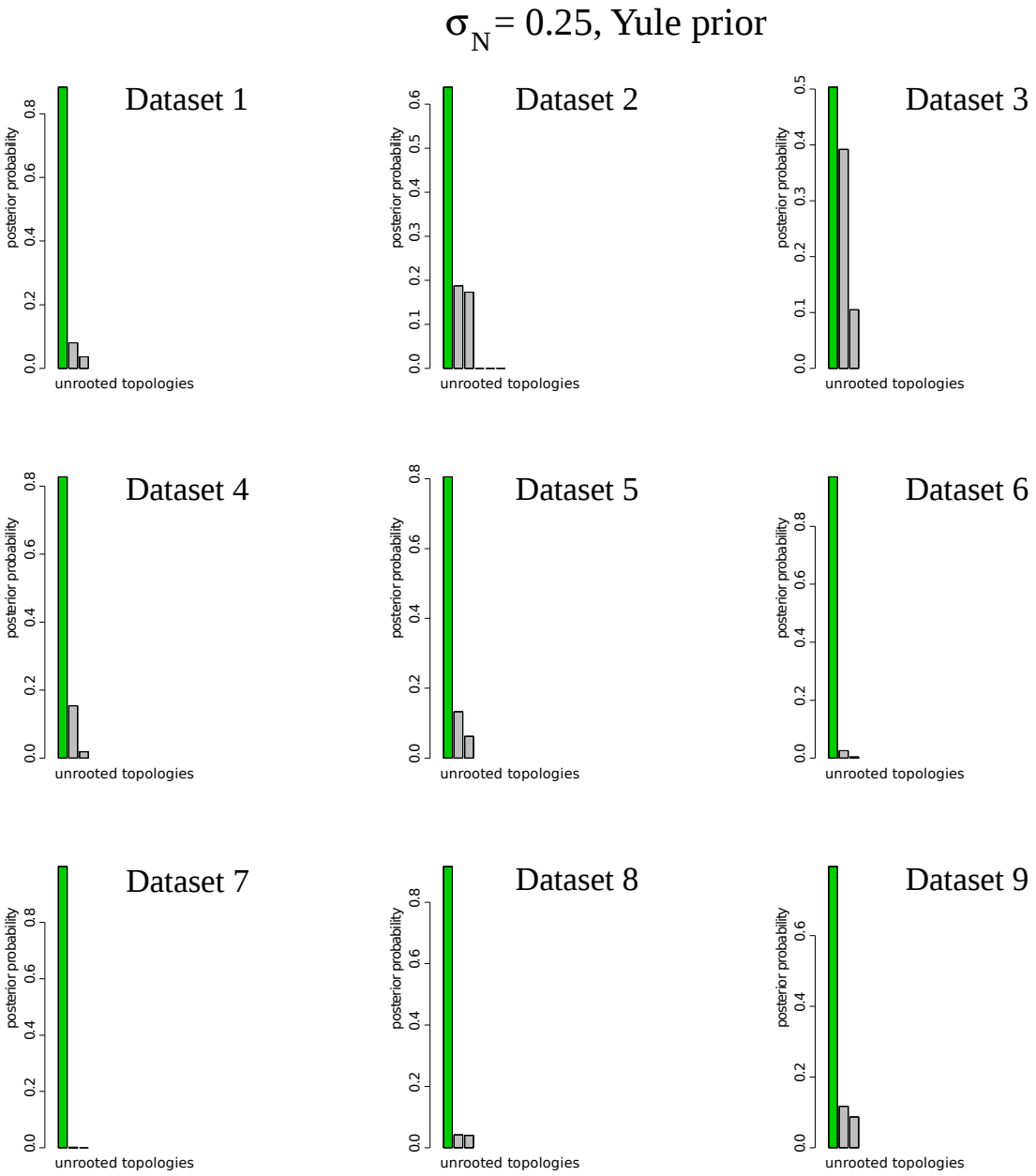
Figure D.1: Posterior distribution of the root splits for different values of σ_N and the Yule prior.



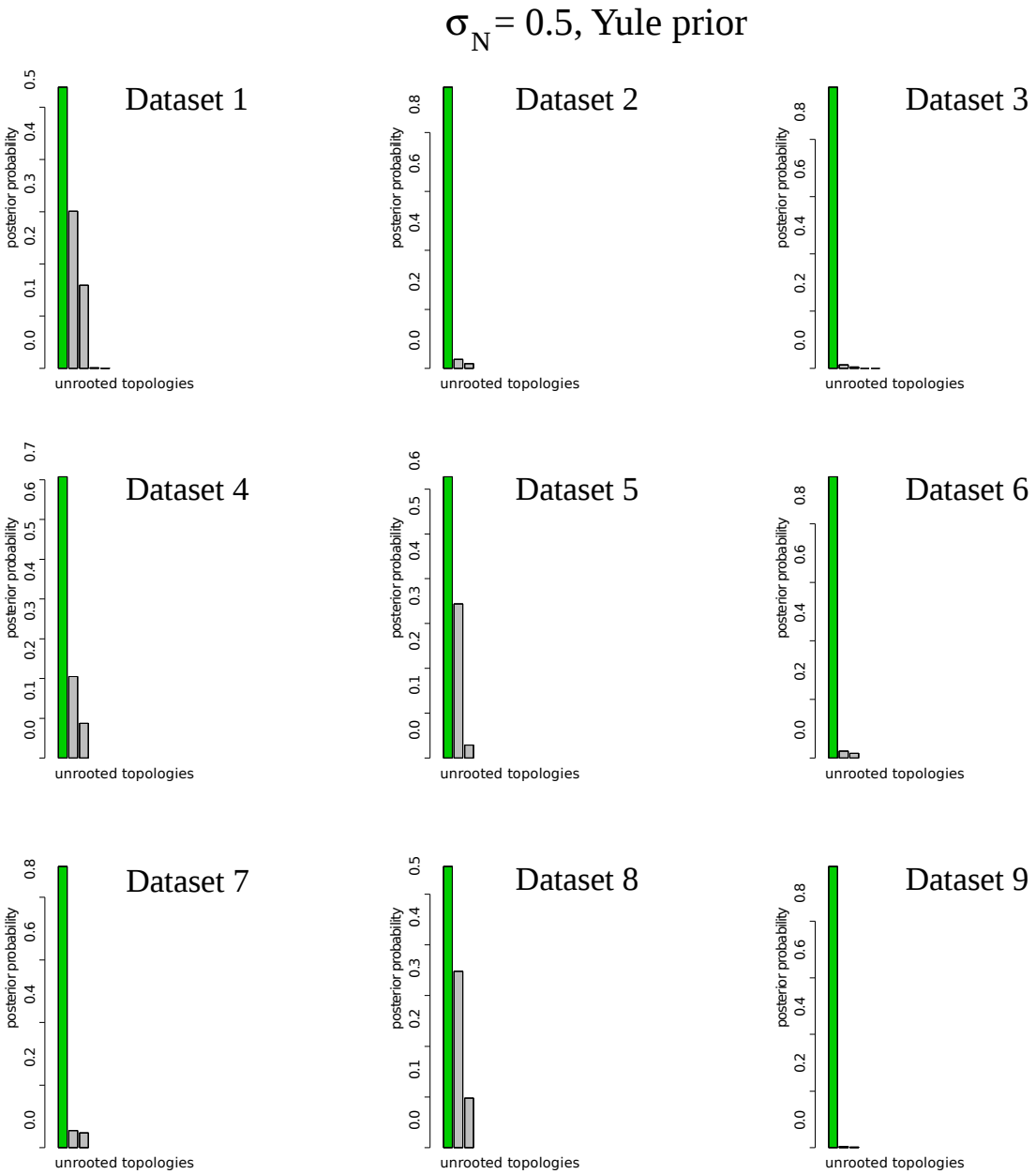
(a) Yule prior, $\sigma_N = 0$.



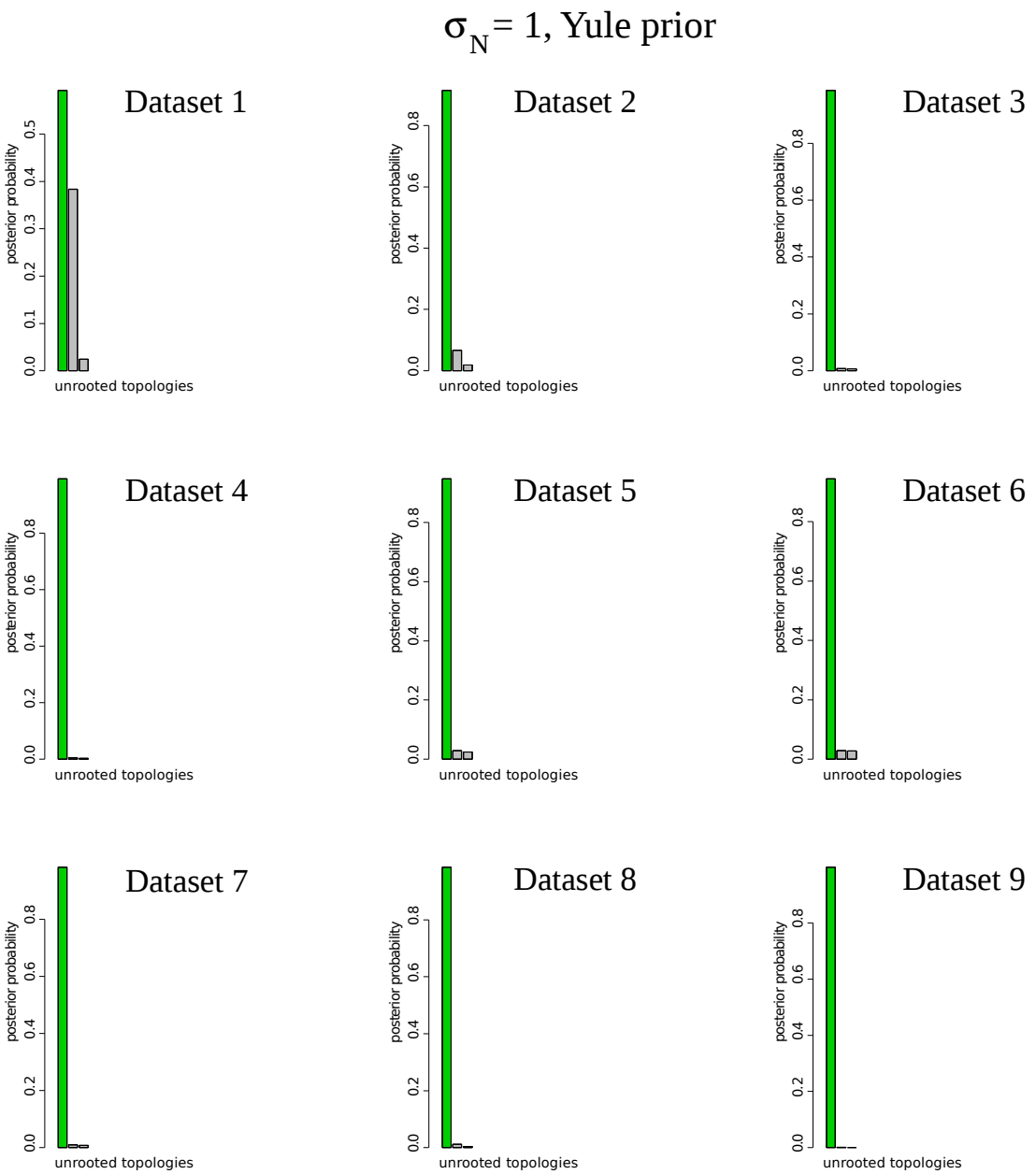
(b) Yule prior, $\sigma_N = 0.1$.



(c) Yule prior, $\sigma_N = 0.25$.



(d) Yule prior, $\sigma_N = 0.5$.

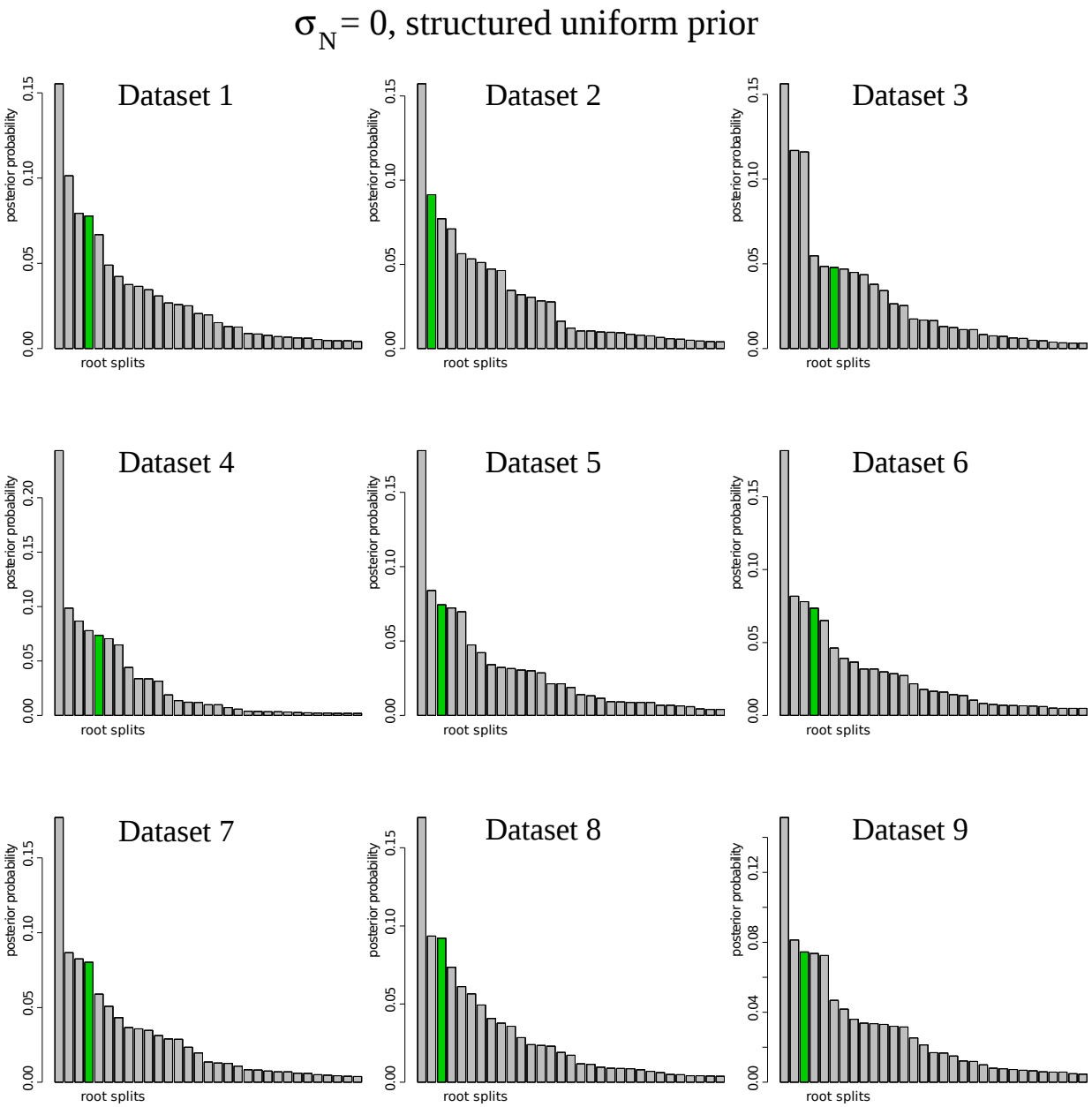


(e) Yule prior, $\sigma_N = 1$.

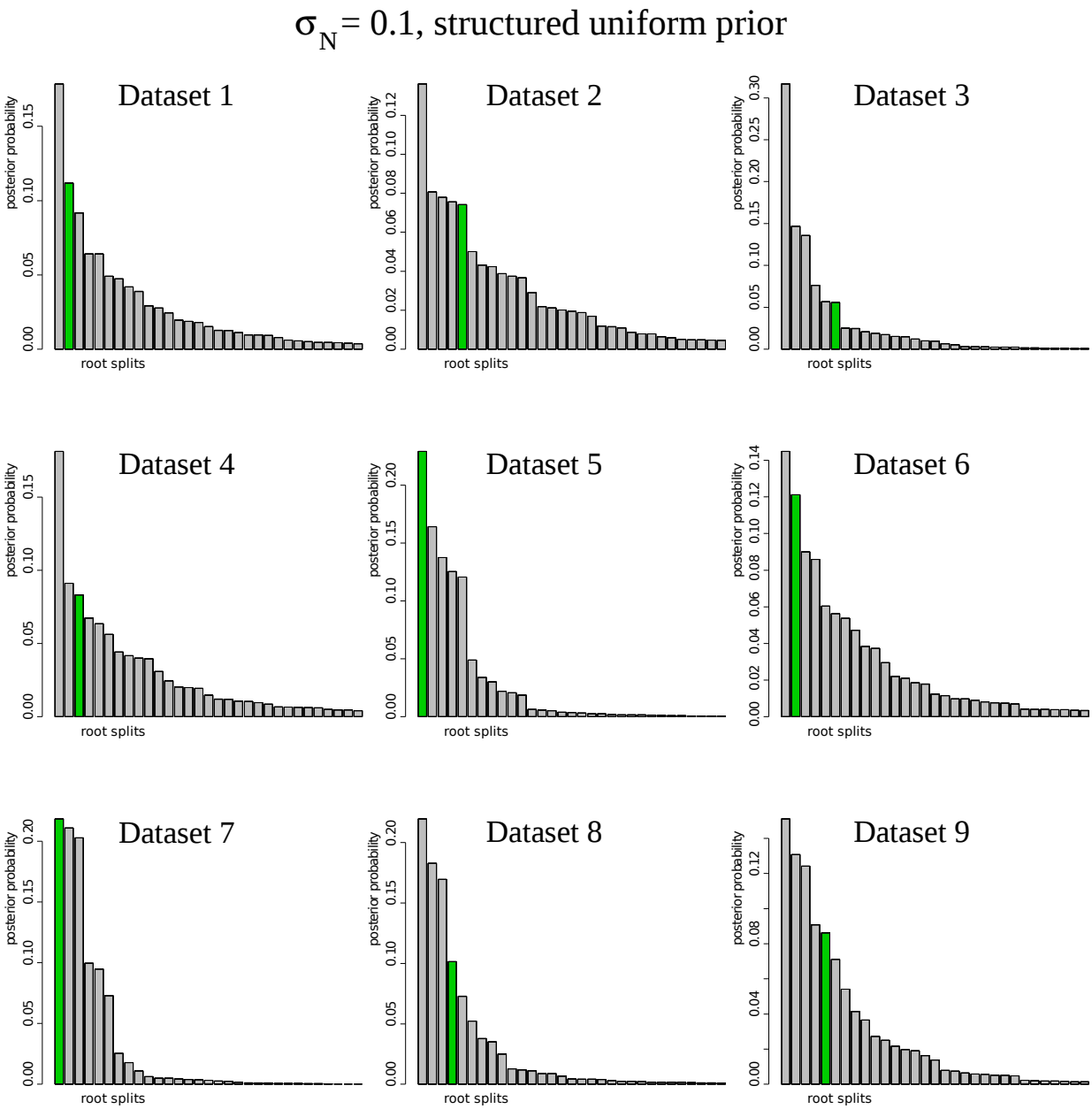
Figure D.2: Posterior distribution of the unrooted topologies for different values of σ_N and the Yule prior.

Appendix E

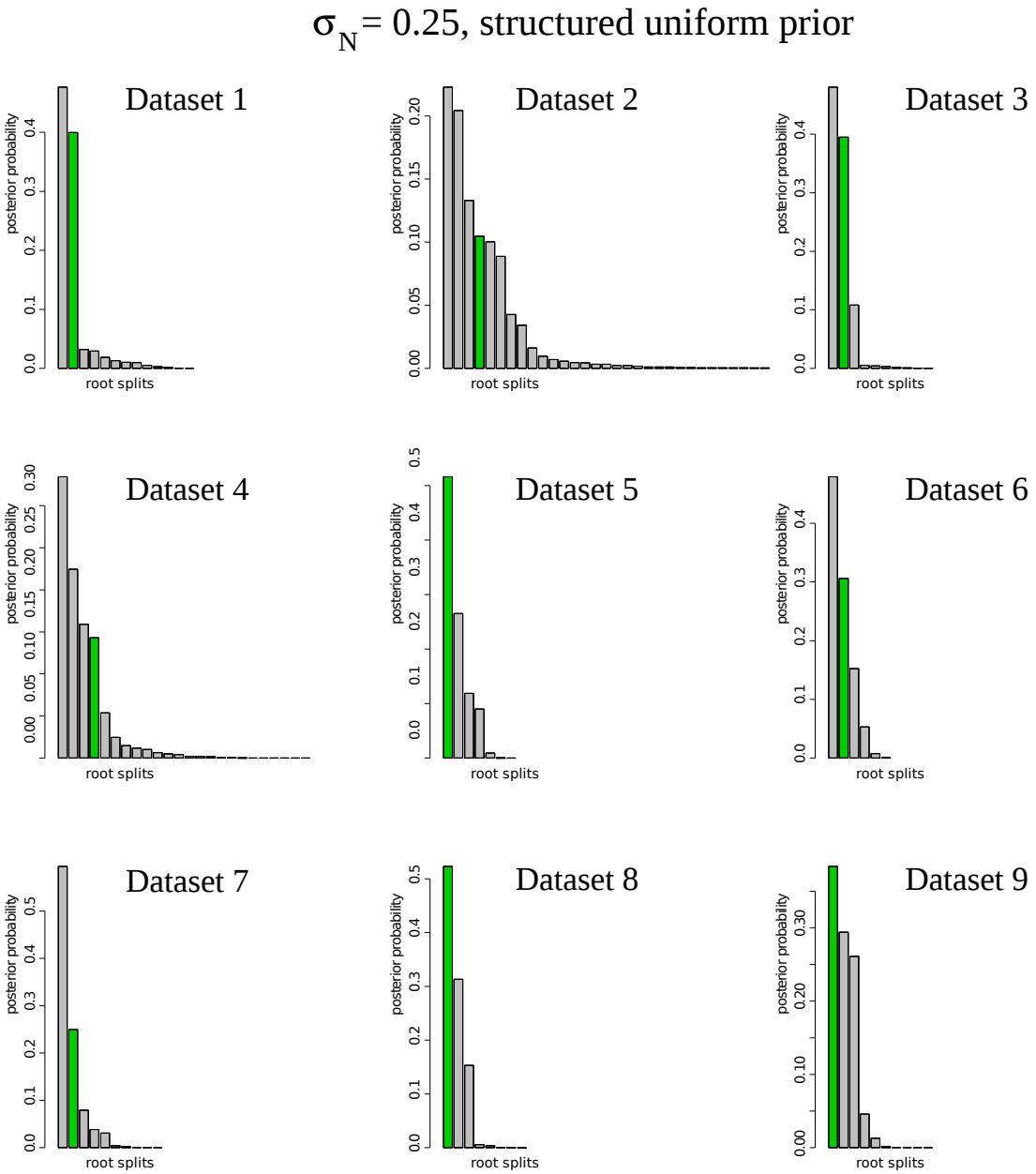
This appendix summarises the results of the first block of simulations for the NR2 model, analysed with the structured uniform prior. Figure E.1 shows the posterior distribution of the root splits for $\sigma_N = 0,\ 0.1,\ 0.25,\ 0.5,\ 1$. Different bars on the plots represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). On each plot the green bar represents the true root split. Figure E.2 shows the posterior distribution of the unrooted topologies for $\sigma_N = 0,\ 0.1,\ 0.25,\ 0.5,\ 1$. Different bars on the plots represent different unrooted topologies (ordered by posterior probabilities). On each plot the green bar represents the true unrooted topology. Each subfigure contains an analysis of nine alignments simulated with a particular value of σ_N.



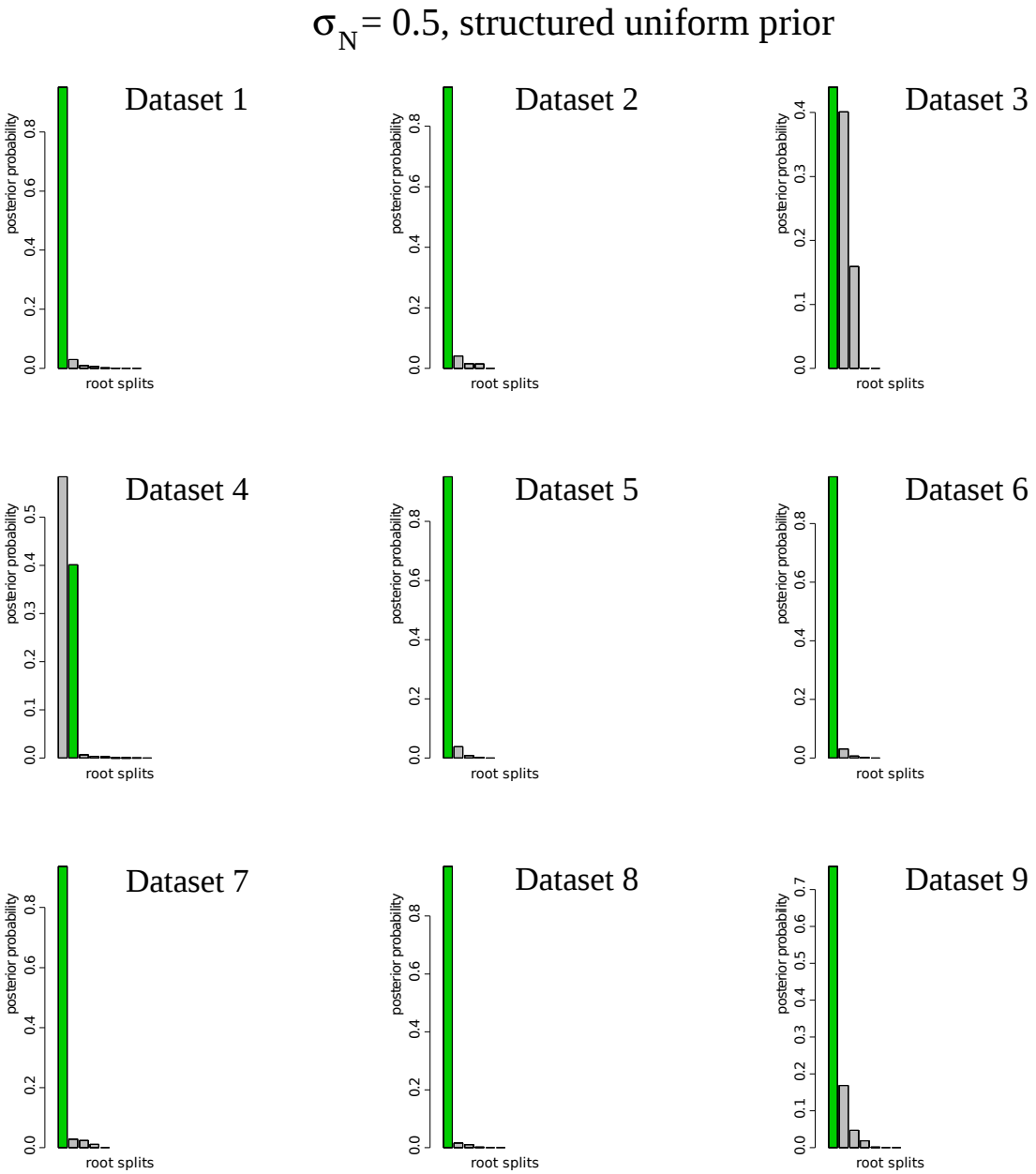
(a) Structured uniform prior, $\sigma_N = 0$.



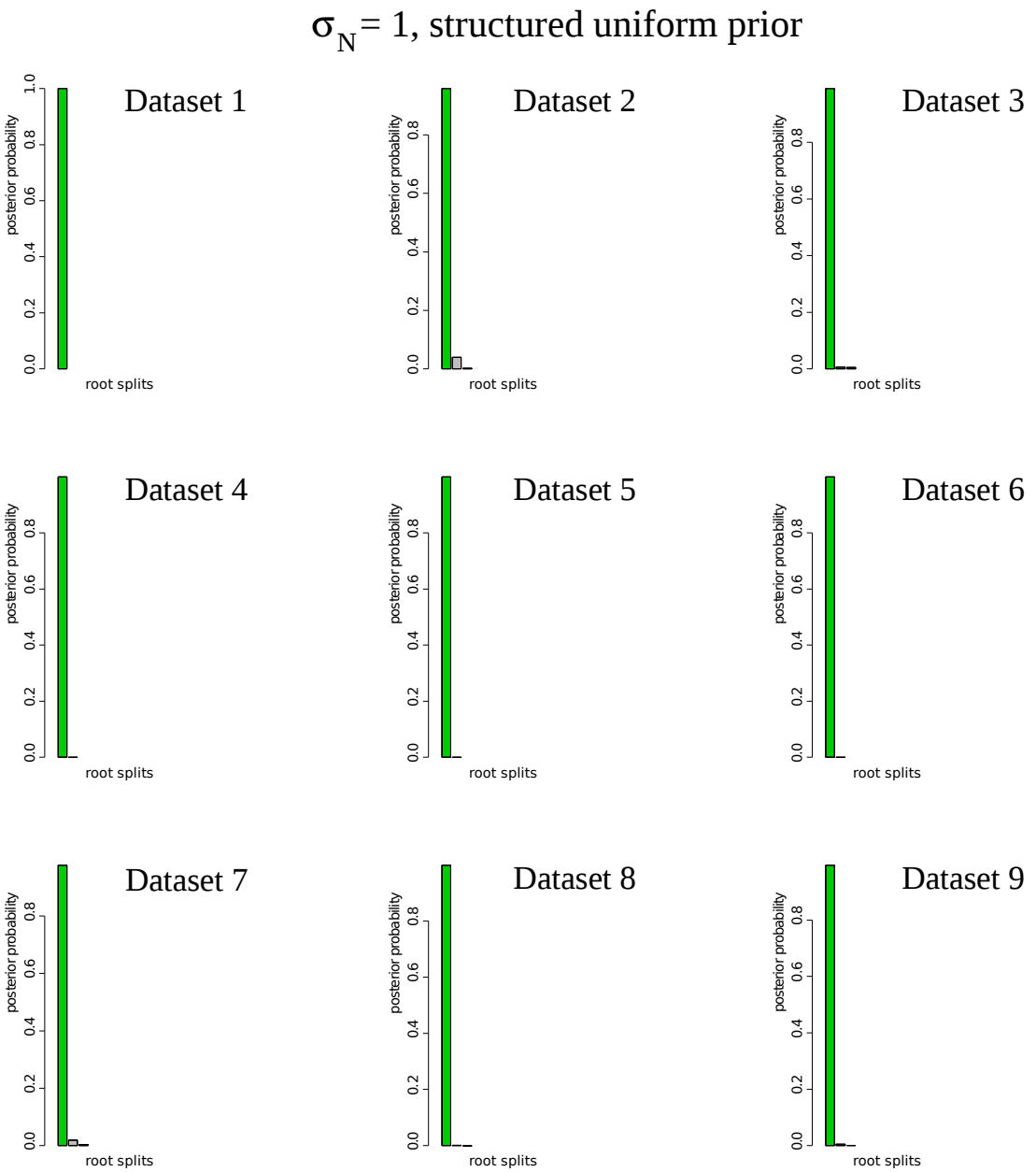
(b) Structured uniform prior, $\sigma_N = 0.1$.



(c) Structured uniform prior, $\sigma_N = 0.25$.

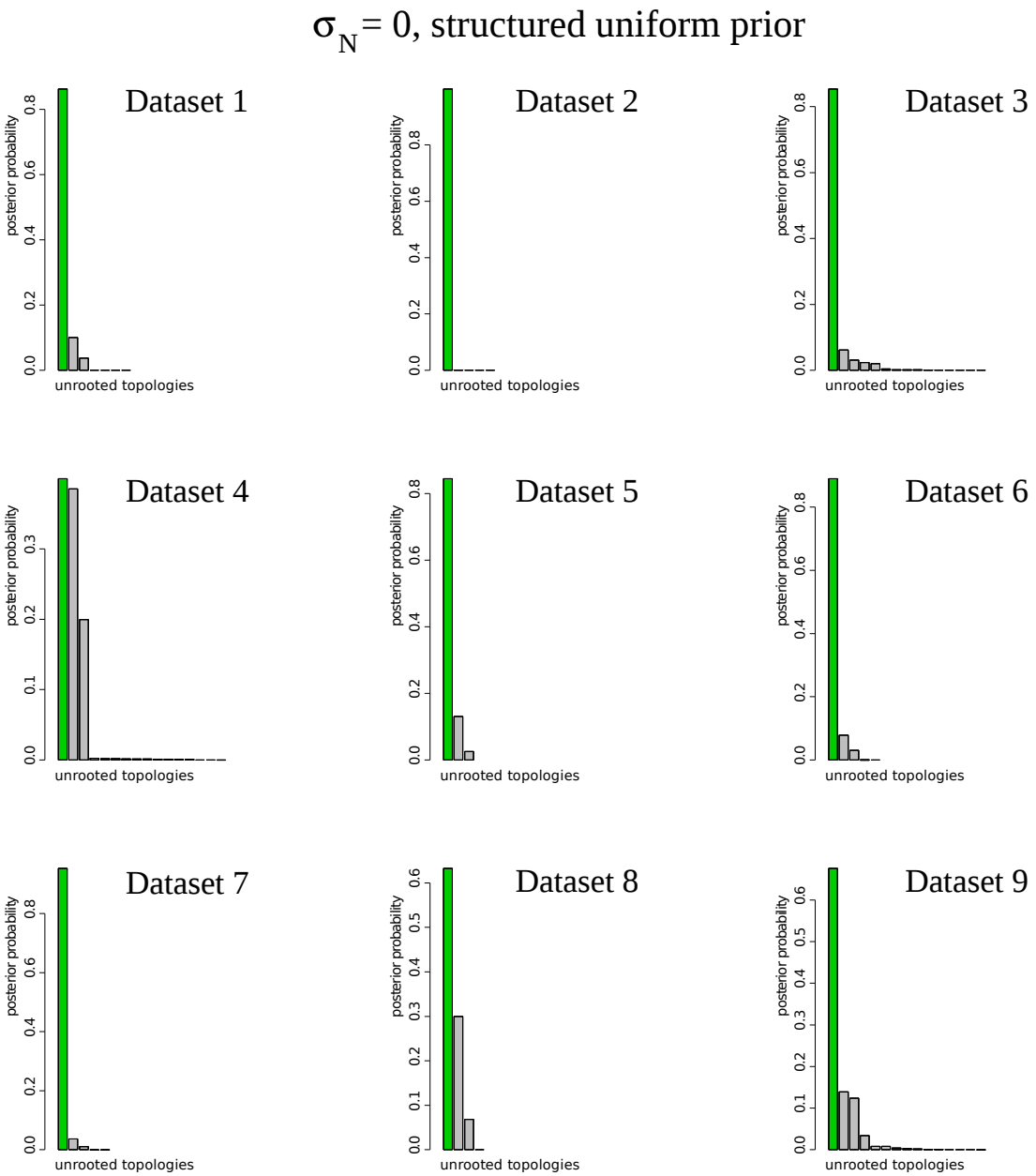


(d) Structured uniform prior, $\sigma_N = 0.5$.

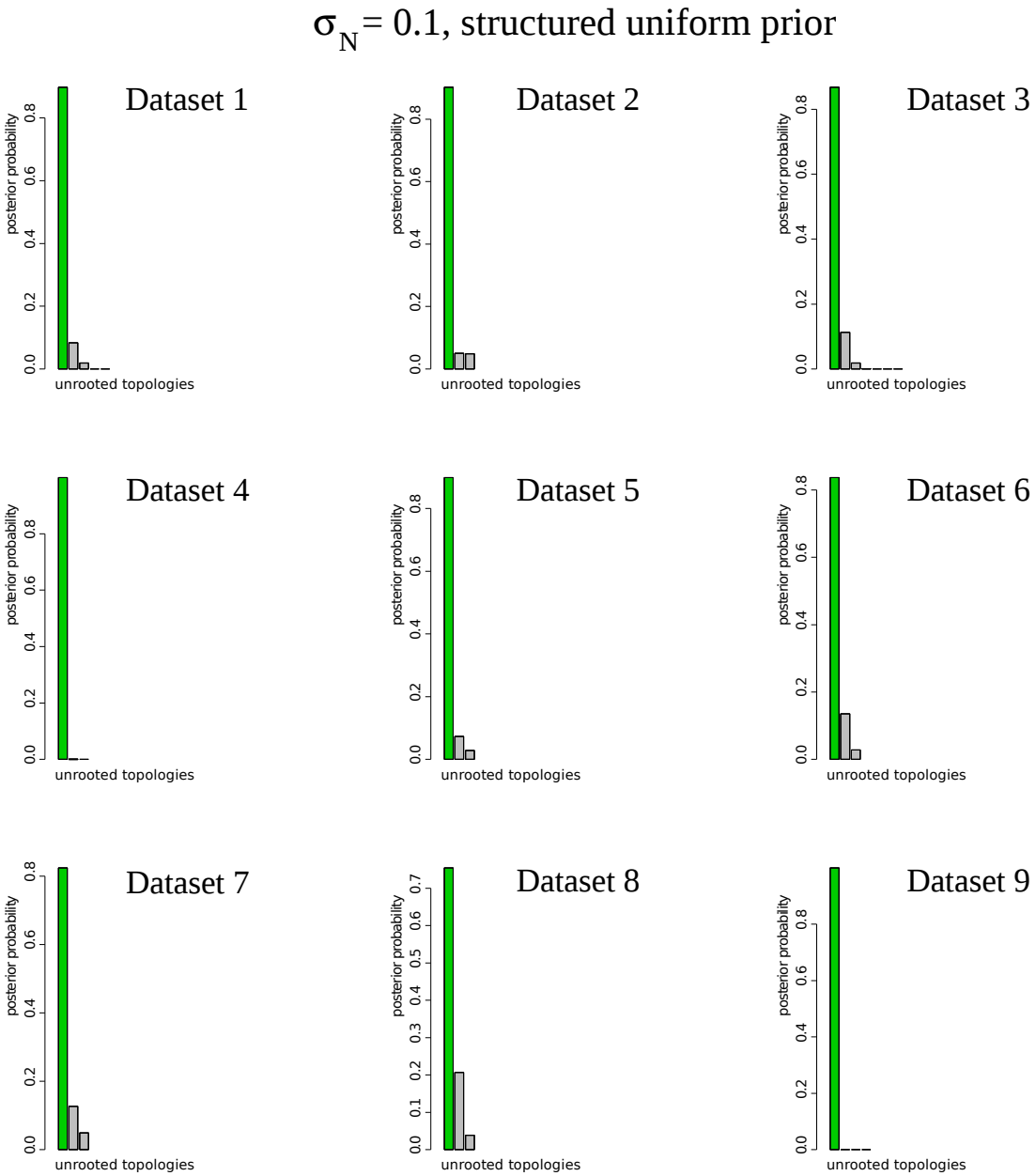


(e) Structured uniform prior, $\sigma_N = 1$.

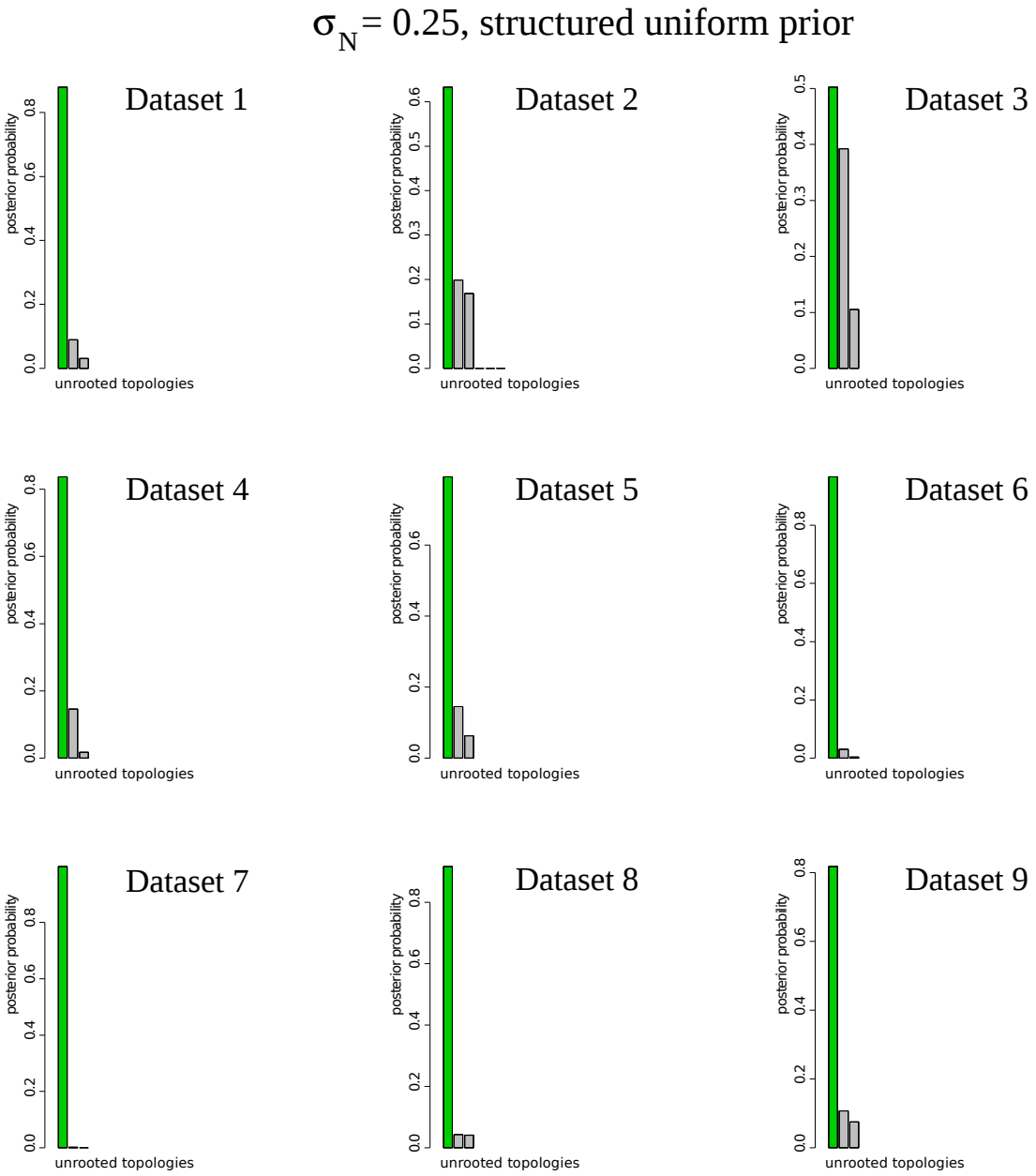
Figure E.1: Posterior distribution of the root splits for different values of σ_N and structured uniform prior.



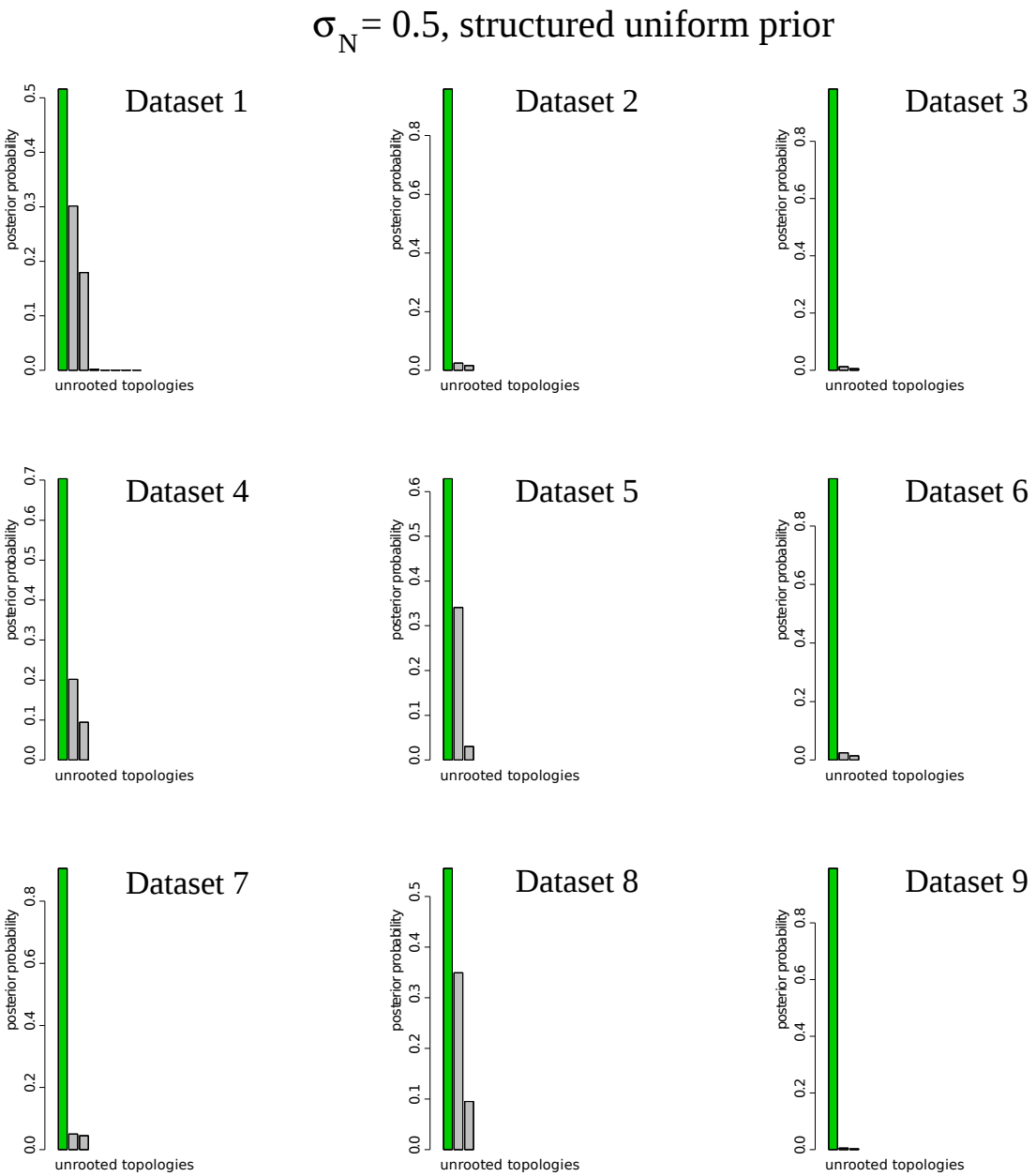
(a) Structured uniform prior, $\sigma_N = 0$.



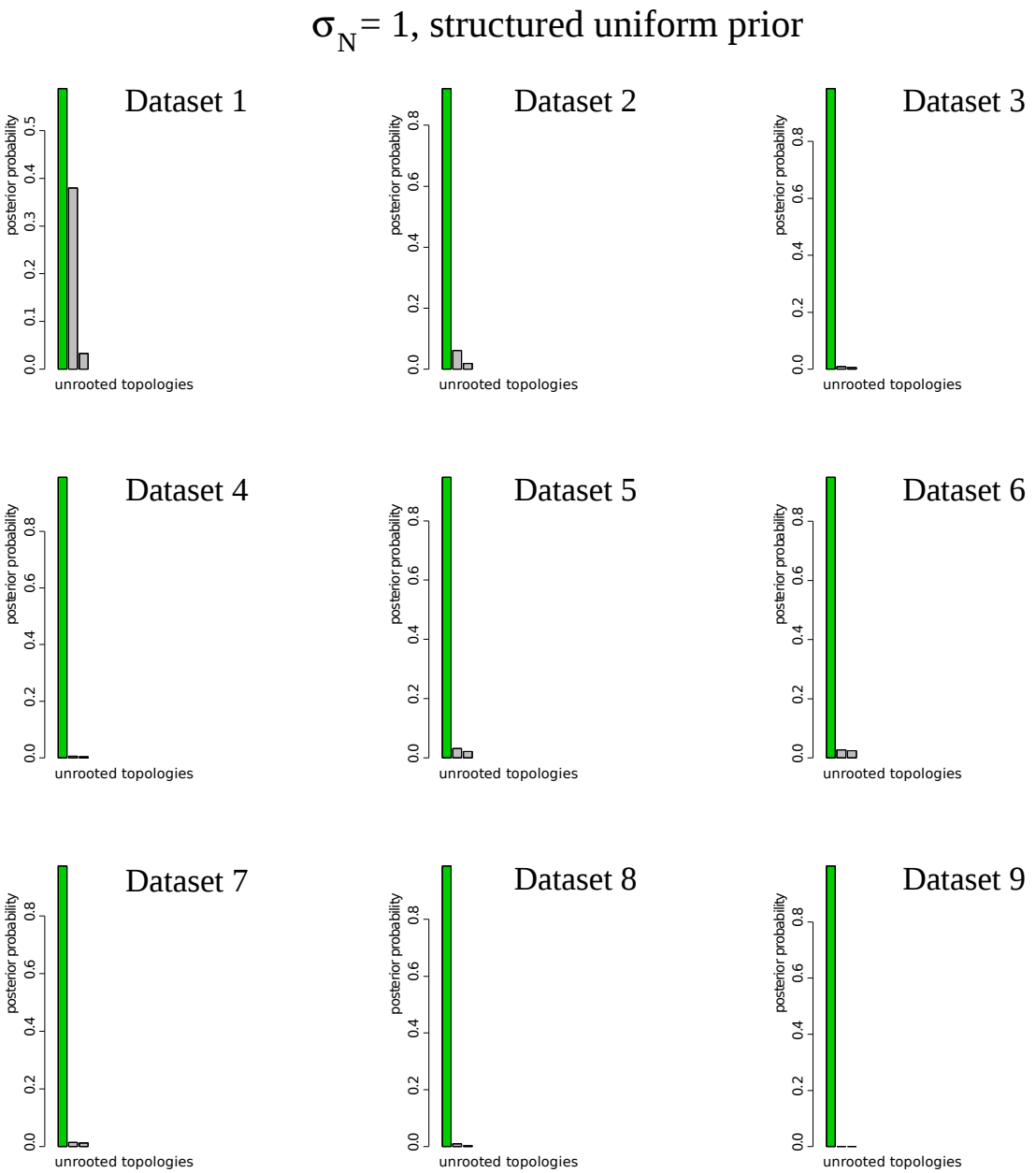
(b) Structured uniform prior, $\sigma_N = 0.1$.



(c) Structured uniform prior, $\sigma_N = 0.25$.



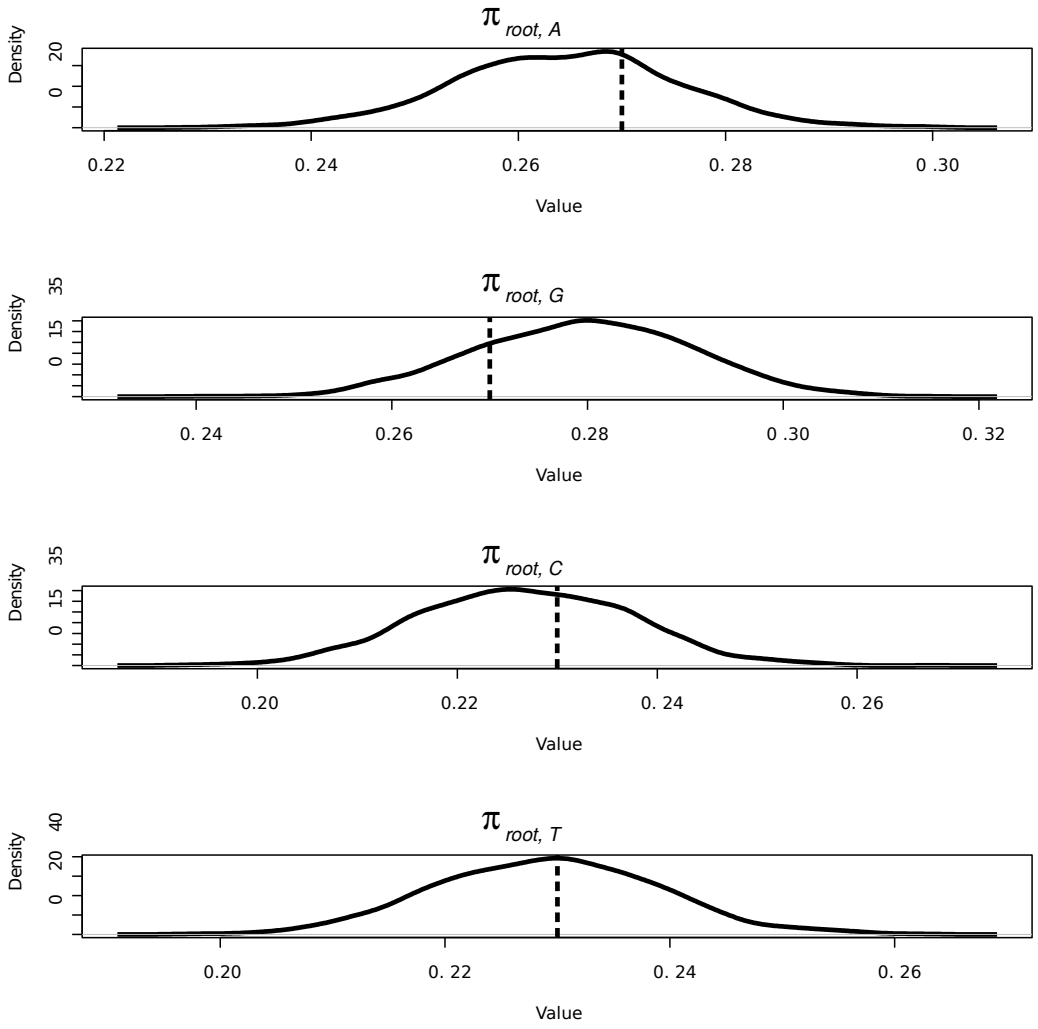
(d) Structured uniform prior, $\sigma_N = 0.5$.



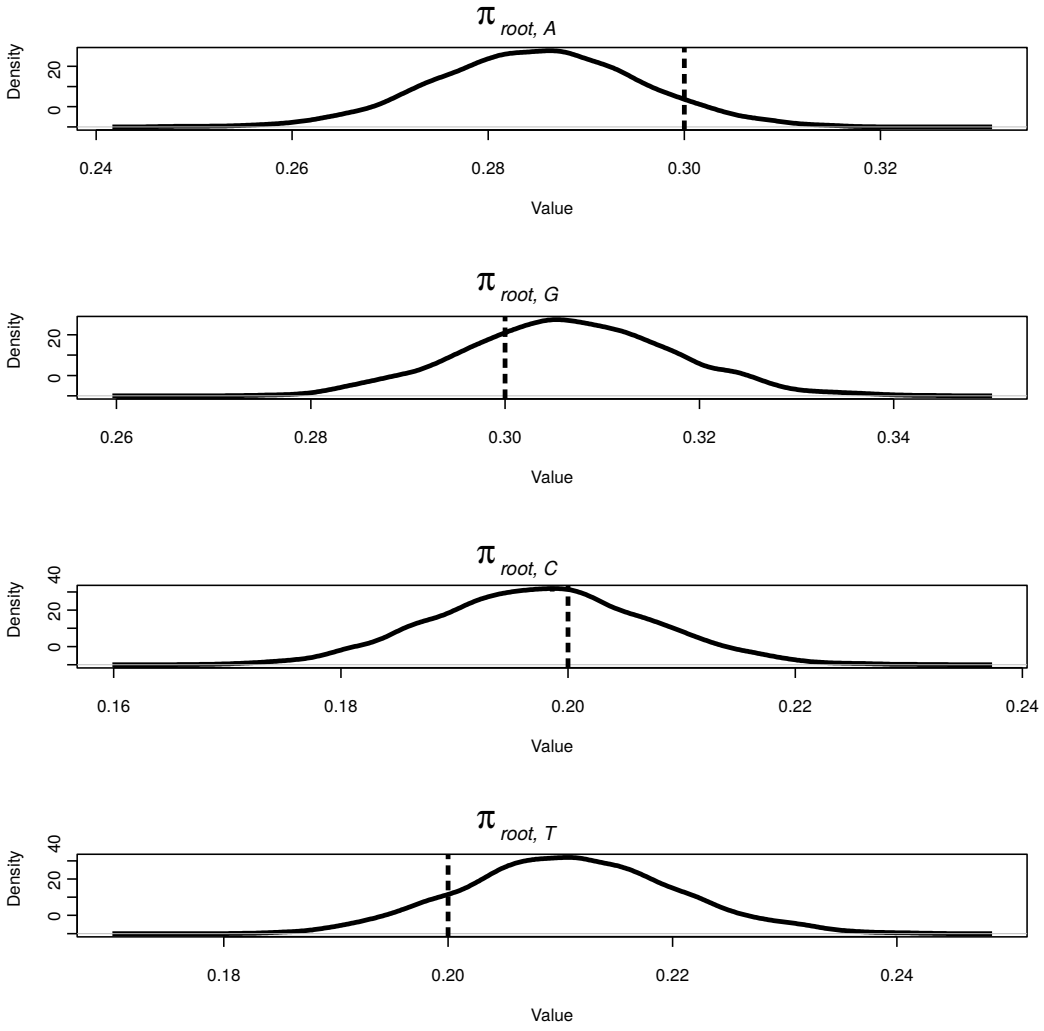
(e) Structured uniform prior, $\sigma_N = 1$.

Figure E.2: Posterior distribution of the unrooted topologies for different values of σ_N and structured uniform prior.

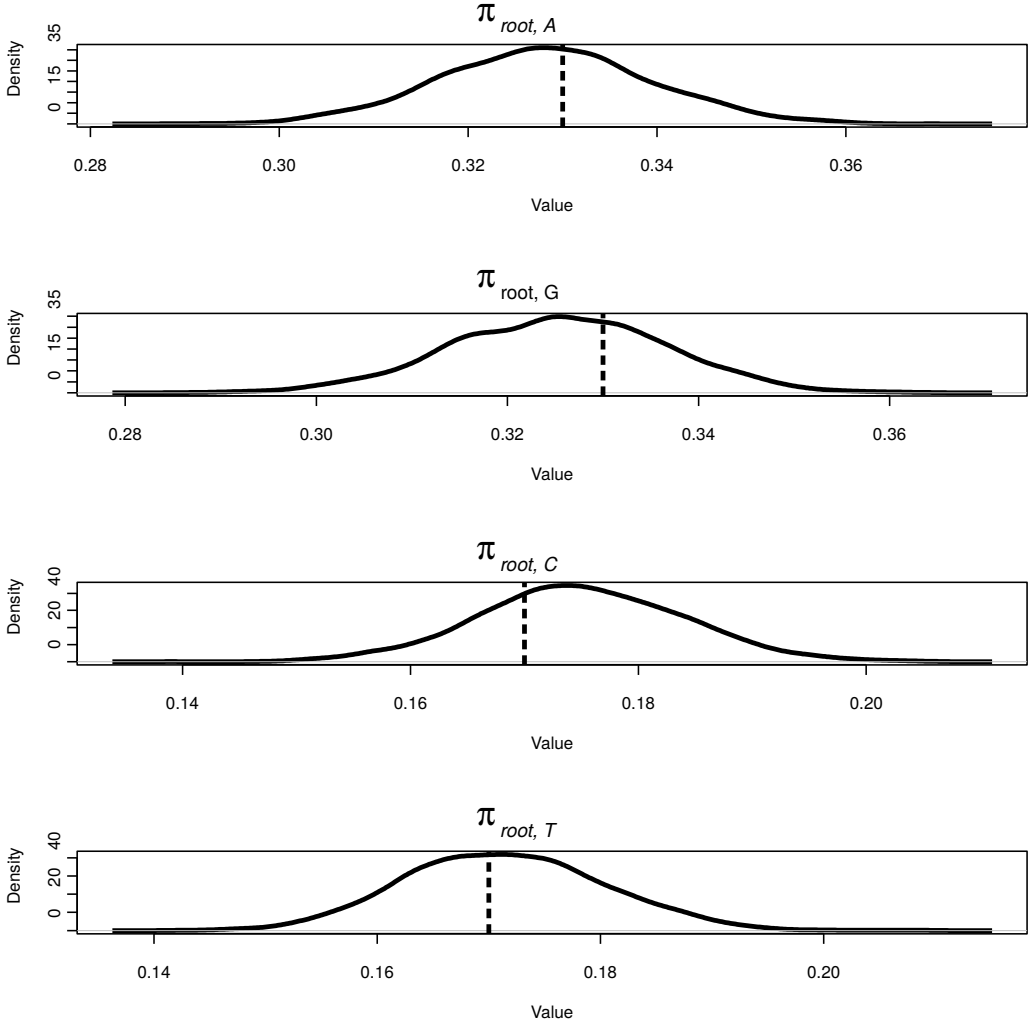
Appendix F



(a) Posterior distribution of $\boldsymbol{\pi}_{root}$ for the dataset with low level of non-stationarity ($\boldsymbol{\pi}_{root}$ = (0.27, 0.27, 0.23, 0.23)).



(b) Posterior distribution of $\boldsymbol{\pi}_{root}$ for the dataset with moderate level of non-stationarity ($\boldsymbol{\pi}_{root}$ = (0.3, 0.3, 0.2, 0.2)).



(c) Posterior distribution of $\boldsymbol{\pi}_{root}$ for the dataset with high level of non-stationarity ($\boldsymbol{\pi}_{root} = (0.33,\ 0.33,\ 0.17,\ 0.17)$).

Figure F.1: Posterior distribution of the composition at the root $\boldsymbol{\pi}_{root}$ for three datasets simulated with different levels of non-stationarity: (a) low level of non-stationarity ($\boldsymbol{\pi}_{root} = (0.27,\ 0.27,\ 0.23,\ 0.23)$); (b) moderate level of non-stationarity ($\boldsymbol{\pi}_{root} = (0.3,\ 0.3,\ 0.2,\ 0.2)$); (c) high level of non-stationarity ($\boldsymbol{\pi}_{root} = (0.33,\ 0.33,\ 0.17,\ 0.17)$). The true values of $\boldsymbol{\pi}_{root}$ are shown with dashed vertical lines.

Appendix G

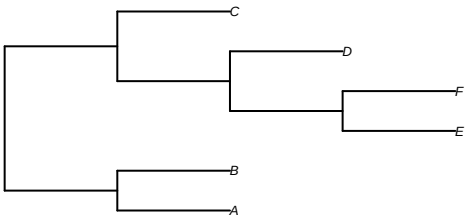
The root on the majority rule consensus tree and the mode of the posterior distribution for root splits are different point summaries of the posterior distribution for root positions. Both can be approximated from posterior samples of rooted topologies but they need not coincide. For example, suppose the posterior output comprises the following five trees:

Tree 1:	((A:1,B:1):1,(((E:1,F:1):1,D:1):1,C:1):1);
Tree 2:	(((A:1,B:1):1,C:1):1,((E:1,F:1):1,D:1):1);
Tree 3:	((((A:1,B:1):1,C:1):1,D:1):1,(E:1,F:1):1);
Tree 4:	(((((A:1,B:1):1,C:1):1,D:1):1,E:1):1,F:1);
Tree 5:	((A:1,B:1):1,(((E:1,F:1):1,D:1):1,C:1):1);

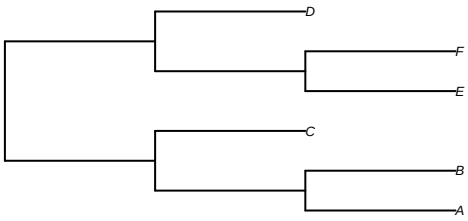
The clade (A, B) appears on all the trees, and so is included in the consensus tree with probability one. Similarly, the clade (A, B, C) appears on three trees (Tree 2, Tree 3 and Tree 4), and so appears in the consensus tree with support 0.6. Continuing in this fashion, the consensus tree is completed by incorporating the clades (E, F) and (D, E, F) which appear with support 0.8 and 0.6 respectively. Hence, the root position on the consensus tree (shown in the Figure below) separates the taxa A, B, C from D, E, F. On the other hand, the posterior for root splits is given by:

Root split	**Count**	**Probability**
(A, B) : (C, D, E, F)	2	0.4
(A, B, C) : (D, E, F)	1	0.2
(E, F) : (A, B, C, D)	1	0.2
(F) : (A, B, C, D, E)	1	0.2

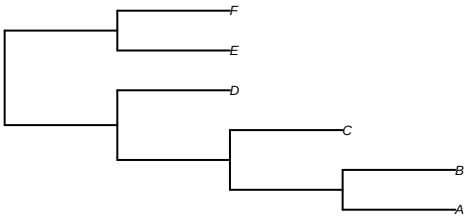
Thus the posterior modal root split is (A, B) : (C, D, E, F) which does not match the root split (A, B, C) : (D, E, F) on the consensus tree.



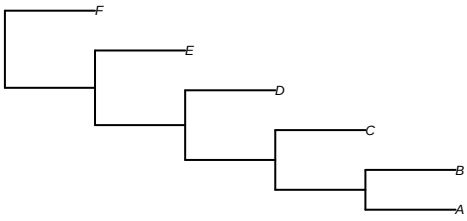
Tree 1



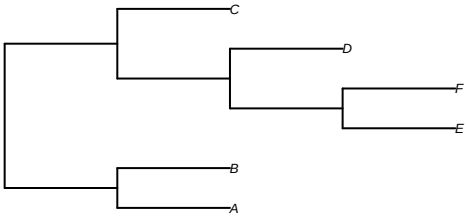
Tree 2



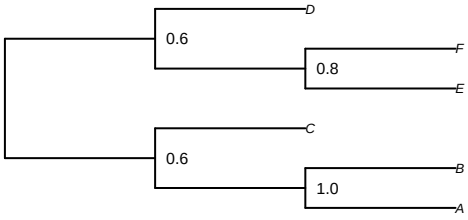
Tree 3



Tree 4



Tree 5



Consensus tree

Bibliography

Abby, S. S., Tannier, E., Gouy, M. & Daubin, V. 2012 Lateral gene transfer as a support for the tree of life. *Proc. Natl. Acad. Sci. USA* **109**, 4962–4967.

Alfaro, M. E. & Holder, M. T. 2006 The posterior and the prior in Bayesian phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **37**, 19–42.

Allen, B. L. & Steeel, M. 2001 Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics* **5**, 1–15.

Baldauf, S. L. 1996 The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* **93**, 7749–7754.

Bergsten, J. 2005 A review of long-branch attraction. *Cladistic* **21**, 163–193.

Blanquart, S. & Lartillot, N. 2006 A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23(11)**, 2058–2071.

Bosquet, J., Strauss, S. H., Doeksen, A. H. & Price, R. A. 1992 Extensive variation in evolutionary rate of rbcL gene sequences among seed plants. *Proc. Natl. Acad. Sci. USA* **89**, 7844–7848.

Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E. & Daubin, V. 2013 Genome-scale coestimation of species and gene trees. *Genome Research* **23**, 323–330.

Brinkmann, H. & Philippe, H. 1999 Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* **16**, 817–825.

Bromham, L., Rambaut, A. & Harvey, P. H. 1996 Determinants of rate variation in mammalian DNA sequence evolution. *J. Mol. Evol.* **43**, 610–621.

Brooks, S. 1998 Markov chain monte carlo method and its application. *The Statistician* **47**, 69–100.

Brown, J. R. & Doolittle, W. F. 1995 Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* **92**, 2441–2445.

Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* **18**, 225–239.

Byrne, K. P. & Wolfe, K. H. 2005 The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research* **15**, 1456–1461.

Cannarozzi, G., Schneider, A. & Gonnet, G. 2007 A phylogenomic study of human, dog, and mouse. *PLoS Computational Biology* **3**, 9–14.

Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. 2009 trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.

Cavalier-Smith, T. 2006 Rooting the tree of life by transition analyses. *Biology Direct* **1**, 19–19.

Chib, S. & Greenberg, E. 1995 Understanding the metropolis-hastings algorithm. *The American Statistician* **49**, 327–335.

Cowles, M. K. & Carlin, B. P. 1996 Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* **91**, 883–904.

Cox, C. J., Foster, P. J., Hirt, R. P., Harris, S. R. & Embley, T. M. 2008 The archaebacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. USA* **51**, 20356–20361.

Curtis, S. E. & Clegg, M. T. 1984 Molecular evolution of chloroplast DNA sequences. *Mol. Biol. Evol.* **4**, 291–301.

Dagan, T., Roettger, M., Bryant, D. & Martin, W. 2010 Genome networks root the tree of life between prokaryotic domains. *Genome Biol. Evol.* **2**, 379–392.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. 1989 A model of evolutionary change in proteins. *Atlas of Protein Sequences and Structure* **5**, 345–352.

Dutheil, J. & Boussau, B. 2008 Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol. Biol.* **28**, 255.

Edgar, R. C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797.

Embley, T. M., van der Giezen, M., Horner, D. S., Dyal, P. L. & Foster, P. 2003 Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Phil. Trans. R. Soc. Lond. B* **358**, 191–203.

Embley, T. M. & Martin, W. 2006 Eukaryotic evolution, changes and challenges. *Nature* **440**, 623–630.

Farris, J. S. 1972 Estimating phylogenetic trees from distance matrices. *The American Naturalist* **106**, 645–668.

Felsenstein, J. 1973 Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Biol.* **22**, 240–249.

Felsenstein, J. 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.* **27**, 401–410.

Felsenstein, J. 1981 Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.

Foster, P. G. 2004 Modeling compositional heterogeneity. *Syst. Biol.* **53(3)**, 485–495.

Galtier, N. & Gouy, M. 1998 Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **15(70**, 871–879.

Gelfand, A. E. & Smith, F. M. 1990 Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. & Rubin, D. 1992 Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

Germain, S. E. 2010 Bayesian spatiotemporal modelling of rainfall through nonhomogeneous hidden markov models. *PhD Thesis* **5.6.1**, 188–189.

Geweke, J. 1992 Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.

Geyer, C. J. 1992 Practical markov chain monte carlo. *Statistical Science* **7**, 473–511.

Geyer, C. J. 2011 *Handbook of Markov Chain Monte Carlo in Introduction to Markov Chain Monte Carlo*. Springer, New York.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. 1996 *Markov Chain Monte Carlo in Practice*. London, Chapman & Hall.

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Data, T. & Oshima, T. 1989 Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. USA* **86**, 6661–6665.

Gojobori, T., Li, W.-H. & Graur, D. 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**, 360–369.

Gouy, M. & Li, W. H. 1989 Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. *Nature* **339**, 145–147.

Guy, L. & Ettema, T. J. G. 2011 The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends in Microbiology* **19**, 580–587.

Hasegawa, M., Kishino, H. & Yano, T. 1985 Dating of human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.

Hashimoto, T. & Hasegawa, M. 1996 Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G. *Adv. Biophys.* **32**, 73–120.

Hastings, W. K. 1970 Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**, 97–109.

Heaps, S. E., Nye, T. M. W., Boys, R. J., Williams, T. A. & Embley, T. M. 2014 Bayesian modelling of compositional heterogeneity in molecular phylogenetics. *Stat. Appl. Genet. Mol. Biol.* **1**, 1–21.

Hedtke, S. M., Townsend, T. M. & Hillis, D. M. 2006 Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* **55**, 522–529.

Hess, P. N. & Russo, C. A. D. M. 2013 An empirical test of the midpoint rooting method. *Biological Journal of the Linnean Society* **92**, 669–674.

Holland, B. R., Penny, D. & Hendy, M. D. 2003 Outgroup misplacement and phylogenetic inaccuracy under a molecular clock —a simulation study. *Syst. Biol.* **52**, 229–238.

Hrdy, I., Hirt, R. P., , Dolezal, P., Bardonová, L., Forster, P. G., Tachezy, J. & Embley, T. M. 2004 Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **423**, 618–622.

http://ygob.ucd.ie 2015 Yeast Gene Order Browser. `http://ygob.ucd.ie//`, [Online; accessed 1-Jan-2015].

Huelsenbeck, J. P., Bollback, J. P. & Levine, A. M. 2002 Inferring the root of a phylogenetic tree. *Syst. Biol.* **51(1)**, 32–43.

Huelsenbeck, J. P. & Ronquist, F. 2001 MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**, 754–755.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. 1989 Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* **86**, 9355–9359.

Jayaswal, V., Ababneh, F., Jermiin, L. S. & Robinson, J. 2011 Reducing model complexity of the general Markov model of evolution. *Mol. Biol. Evol.* **28(11)**, 3045–3059.

Jayaswal, V., Robinson, J. & Jermiin, L. S. 2007 Estimation of phylogeny and invariant sites using the general Markov model of nucleotide sequence evolution. *Syst. Biol.* **56**, 155–162.

Jayaswal, V., Wong, T. K. F., Robinson, J., L.Poladian & Jermiin, L. S. 2014 Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Syst. Biol.* **63**, 726–742.

Jones, D. T., Taylor, W. R. & Thornton, J. M. 1992 The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275–282.

Jukes, T. H. & Cantor, C. R. 1969 *Evolution of Protein Molecules*. New York: Academic Press.

Kellis, M., Birren, B. W. & Lander, E. S. 2004 Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–624.

Kelly, S., Wickstead, B. & Gull, K. 2011 Archaeal phylogenomics provides evidence in support of a methanogenic origin of the archaea and a thaumarchaeal origin for the eukaryotes. *Proc. R. Soc. Lond. B.* **278**, 1009–1018.

Kimura, M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120.

Kumar, S. 2005 Molecular clocks: four decades of evolution. *Nature Reviews Genetics* **6**, 654–662.

Lake, J. A. 1988 Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331**, 184–186.

Lake, J. A., Henderson, E., Oakes, M. & Clark, M. W. 1984 Eocytes: A new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl. Acad. Sci. USA* **81**, 3786–3790.

Lake, J. A., Herbold, C. W., Rivera, M. C., Servin, J. A. & G.Skophammer, R. 2007 Rooting the tree of life using nonubiquitous genes. *Mol. Biol. Evol.* **25**, 130–136.

Lake, J. A., Skophammer, R. G., Herbold, C. W. & Servin, J. A. 2009 Genome beginnings: rooting the tree of life. *Phil. Trans. R. Soc. B* **364**, 2177–2185.

Lartillot, N. & Philippe, H. 2004 A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109.

Lasek-Nesselquist, E. & Gogarten, J. P. 2013 The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol. Biol. Evol.* **69**, 17–38.

Le, S. Q. & Gascuel, O. 2008 An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320.

Lopez, P., Casane, D. & Philippe, H. 2002 Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**, 1–7.

Loza-Reyes, E., Hurn, M. A. & Robinson, A. 2014 Classification of molecular sequence data using bayesian phylogenetic mixture models. *Computational Statistics and Data Analysis* **75**, 81–95.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953 Equations of state calculations by fast computing machines. *J. Chem. Phys* **21**, 1087–1091.

Mooers, A. O. & Harvey, P. H. 1994 Metabolic rate, generation time and the rate of molecular evolution in birds. *Mol. Phylogenet. Evol.* **3**, 344–350.

Penny, D. 1976 Criteria for optimising phylogenetic trees and the problem of determining the root of a tree. *Mol. Biol. Evol.* **8**, 95–116.

Penny, D., McComish, B. J., Charleston, M. A. & Hendy, M. D. 2001 Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J. Mol. Evol.* **53**, 711–723.

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., M.Moreira, M. A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P. C., Silva, A., O'Brien, S. J. & econ Slattery, J. 2011 A molecular phylogeny of living primates. *PLoS Genetics* **7**, e1001342.

Philippe, H. & Forterre, P. 1999 The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* **49**, 509–523.

Purvis, A. 1995 A composite estimate of primate phylogeny. *Philosophical Transactions: Biological Sciences* **348**, 405–421.

Renner, S. 2008 Rooting and dating maples (*Acer*) with an uncorrelated-rates molecular clock: implications for North American/Asian disjunctions. *Syst. Biol.* **5**, 795–808.

Rivera, M. C. & Lake, J. A. 1992 Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257**, 74–76.

Roberts, G. O., Gelman, A. & Gilks, W. R. 1997 Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability* **7**, 110–120.

Sanderson, M. T. & Shaffer, H. B. 2002 Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics* **33**, 49–72.

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M. & Wolfe, K. H. 2007 Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc. Natl. Acad. Sci. USA* **104**, 8397–8402.

Sinclair, A. & Jerrum, M. 1989 Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and computation* **82**, 93–133.

Skophammer, R. G., Servin, J. A., Herbold, C. W. & Lake, J. A. 2007 Evidence for a gram-positive, eubacterial root of the tree of life. *Mol. Biol. Evol.* **24**, 1761–1768.

Spang, A., Saw, J. H., L.Jørgensen, S., Niedzwiedzka, K. Z., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L. & Ettema, T. J. G. 2015 Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179.

Squartini, F. & Arndt, P. F. 2008 Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol. Biol. Evol.* **25**, 2525–2535.

Steel, M. & McKenzie, A. 2001 Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences* **170**, 91–112.

Susko, E. & Roger, A. J. 2007 On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150.

Tamura, K. & Nei, M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.

Tarrío, R., Rodríguez-Trelles, F. & Ayala, F. J. 2000 Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the *Drosophila saltans* and *willistoni* groups, a case study. *Mol. Phylogenet. Evol.* **16(3)**, 344–349.

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

Tierney, L. 1994 Markov chains for exploring posterior distributions. *The Annals of Statistics* **22**, 1701–1762.

Whelan, S. & Goldman, N. 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.

Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. 2013 An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236.

Williams, T. A., Foster, P. G., Nye, T. M. W., Cox, C. J. & Embley, T. M. 2012 A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. R. Soc. B* **279**, 4870–4879.

Williams, T. A., Heaps, S. E., Cherlin, S., Nye, T. M. W., Boys, R. J. & Embley, T. M. 2015 New substitution models for rooting phylogenetic trees. *Phil. Trans. R. Soc. B* **370**.

Woese, C. R. 1990 Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**, 4576–4579.

Woese, C. R. & Fox, G. E. 1977 Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **24**, 5088–5090.

Wolfe, K. H. & Shields, D. C. 1997 Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713.

Yang, Z. 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**, 1396–1401.

Yang, Z. 1994 Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**, 306–314.

Yang, Z. 2006 *Computational molecular evolution*. Oxford University Press.

Yang, Z. 2014 *Molecular Evolution: A statistical approach*. Oxford University Press.

Yang, Z. & Rannala, B. 2005 Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* **53**, 455–470.

Yang, Z. & Roberts, D. 1995 On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.* **12**, 451–458.

Yap, V. B. & Speed, T. 2005 Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evol. Biol.* **5**, 2.

Zhaxybayeva, O., Lapierre, P. & Gogarten, J. P. 2005 Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* **227**, 53–64.