Bayesian Phylogenetic Modelling of Lateral Gene Transfers

Rute Gomes Velosa Vieira

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics & Statistics
Newcastle University
Newcastle upon Tyne
United Kingdom

July 2015

This thesis is dedicated to my daughter Catarina, my husband Manuel and my parents.

Acknowledgements

I wish to express my sense of gratitude to one and all, who directly or indirectly, have contributed to the completion of this thesis. Foremost, I am thankful to God for His guidance and strength. I am highly indebted to my daughter Catarina who has been a great source of motivation and inspiration, to my husband Manuel who has been a constant source of support and encouragement during the challenging moments, to my parents who have taught me to work hard for the things that I aspire to achieve and to my sister and brother for being there for me when needed. I would like also to express my sincere thanks to Prof. Richard Boys and Dr. Tom Nye, for their inspiring and encouraging way to guide me through this journey and their invaluable comments in the writing of this thesis. I am also grateful to the School of Mathematics and Statistics for providing me with all the necessary facilities and human resources that were key in the work developed.

Abstract

Phylogenetic trees represent the evolutionary relationships between a set of species. Inferring these trees from data is particularly challenging sometimes since the transfer of genetic material can occur not only from parents to their offspring but also between organisms via lateral gene transfers (LGTs). Thus, the presence of LGTs means that genes in a genome can each have different evolutionary histories, represented by different gene trees.

A few statistical approaches have been introduced to explore non-vertical evolution through collections of Markov-dependent gene trees. In 2005 Suchard described a Bayesian hierarchical model for joint inference of gene trees and an underlying species tree, where a layer in the model linked gene trees to the species tree via a sequence of unknown lateral gene transfers. In his model LGT was modeled via a random walk in the tree space derived from the subtree prune and regraft (SPR) operator on unrooted trees. However, the use of SPR moves to represent LGT in an unrooted tree is problematic, since the transference of DNA between two organisms implies the contemporaneity of both organisms and therefore it can allow unrealistic LGTs.

This thesis describes a related hierarchical Bayesian phylogenetic model for reconstructing phylogenetic trees which imposes a temporal constraint on LGTs, namely that they can only occur between species which exist concurrently. This is achieved by taking into account possible time orderings of divergence events in trees, without explicitly modelling divergence times. An extended version of the SPR operator is introduced as a more adequate mechanism to represent the LGT effect in a tree. The extended SPR operation respects the time ordering. It additionaly differs from regular SPR as it maintains a 1-to-1 correspondence between points on the species tree and points on each gene tree. Each point on a gene tree represents the existence of a population containing that gene at some point in time. Hierarchical phylogenetic models were used in the reconstruction of each gene tree from its corresponding gene alignment, enabling the pooling of information across genes. In addition to Suchard's approach, we assume variation in the rate of evolution between different sites. The species tree is assumed to be fixed.

A Markov Chain Monte Carlo (MCMC) algorithm was developed to fit the model in a Bayesian framework. A novel MCMC proposal mechanism for jointly proposing the gene tree topology and branch lengths, LGT distance and LGT history has been developed as well as a novel graphical tool to represent LGT history, the LGT Biplot. Our model was applied to simulated and experimental datasets. More specifically we analysed LGT/reassortment presence in the evolution of 2009 Swine-Origin Influenza Type A virus. Future improvements of our model and algorithm should include joint inference of the species tree, improving the computational efficiency of the MCMC algorithm and better consideration of other factors that can cause discordance of gene trees and species trees such as gene loss.

Contents

List of Figures

List of Tables

1

Introduction

Mutation, natural selection and genetic drift were, for many decades, addressed as the principal mechanisms of *evolution*, a slow mutational process that gradually changed the characteristics of species over the course of time. As a single evolutionary lineage splits into two or more genetically independent ones, new and distinct species have their origin in a process denominated *speciation*. These new lineages will continue to evolve and split independently producing a branching pattern of species. Its reconstruction can be attempted by studying inherited species' characteristics such as DNA, proteins or phenotypic characteristics.

The most commonly used approach to describe species evolutionary relationships is named *cladistics*. It groups organisms together based on whether or not they have one or more shared unique characteristics that come from the group's last common ancestor and are not present in more distant ancestors. These groups are identified by sharing unique features that are not present in other species. Therefore, members of the same group are thought to share a common history and are considered to be more closely related (Harvey & Pagel, 1991). Initially, morphological and physiological features of species were commonly used as the characters, but with the development of molecular biology methodologies and the increasing amount of genetic data becoming available, molecular phylogenetics took the lead in reconstructing the history of inheritance of genetic sequence data from contemporary organisms into a tree-like structure, formally referred to as a *phylogenetic tree*. The very nature of DNA allows it to be used as a "document" of evolutionary history. Comparisons of the DNA sequences of various genes between different organisms can tell us a lot about the relationships of organisms that cannot be correctly inferred from morphology.

Phylogenetic trees are usually built by analysing DNA or protein sequence data from a gene that is present in the genomes of all the species under analysis. Such trees form a vital part of many kinds of biological analysis, and have a wide range of applications in comparative genomics (e.g. identifying how genes are gained, lost and rearranged in related species), population biology (e.g. identifying the path of early human migrations), and biomedicine (e.g. tracing infection pathways for HIV and other pathogens). However, although there is an ever-increasing amount of genetic data available, this flood of new data may not lead directly to a commensurate gain in knowledge, and today, as new population genomic data sets are emerging, our skills of analysis and interpretation are partly overwhelmed. Tree construction is a difficult task: it is very demanding computationally, and in addition the resultant trees are subject to a high degree of uncertainty.

Phylogenetic tree reconstruction from genetic sequence data is based on models representing how sequences change over evolutionary time. When assuming that, as time increases from the moment two sequences diverge from their last common ancestor, so does the number of differences between them, estimating a tree seems to be relatively simple. A measure of similarity between two sequences could result, for example, from counting the number of differences between them and those sequences that are most similar could be grouped together. Nevertheless, the simplicity of such an algorithm underestimates the complexity of the phylogenetic inference problem. A simple measure of the genetic differences between sequences is not necessarily a reliable indication of when they diverged because the evolutionary rate may vary over time (Yang, 2006). The way a sequence evolves when governed by genetic drift should be quite different than when it is influenced by selection. Rates of evolution will vary for genes with different functions, or different parts of a gene with different functions. Similarly, silent (synonymous) sites in protein coding regions will evolve faster than replacement (nonsynonymous) sites due to different functional constraints. Thus, different regions of DNA with different functional constraints will evolve at different rates which might cause distantly related sequences to diverge from each other more slowly than is expected, or even become more similar to each other at some residues.

In addition, the statistical model used to build the tree might not be correct – different models might result in different estimated trees. In particular, the process of nucleotide substitution in DNA is known to be substantially more complex and heterogeneous than the processes currently used in most phylogenetic models. Finally, all estimated trees are subject to random statistical variation: gene sequence data comprises one "sample" from the possible sequences given a fixed evolutionary tree. Random sampling can contribute, to some extent, to tree discordance under any model of tree-based Markovian evolution, and Martyn & Steel (2012) have shown that this effect becomes magnified as branches in the tree become very short, or very long. Also, long branch attraction or model 'mis-specification' may contribute to systematic errors when using some tree reconstruction methods (Felsenstein, 2004).

Alternatively, those conflicts might reflect biological processes which originated different evolutionary patterns for different genes in a genome. Citing Rusin *et al.*

(2014), “the evolution of the genome, apart from the mutation process, is an entangled complex of individual and concerted evolutions of genes, their regulations, gene content and arrangement on chromosomes, genetic flows between the genome and intracellular organelles, and so forth. Their evolutionary histories often do not coincide with each other and with patterns of speciation giving rise to a variety of evolutionary events, such as gene duplications, losses, gains, lateral transfers, chromosome rearrangements, and others”. Therefore, phylogenetic trees built with the information on one specific gene (*gene tree*) might differ between genes. Consequently, some gene trees will differ from the *species tree*, which represents the evolutionary history of the species in study, giving rise to what is referred to as *phylogenetic incongruence*. Interestingly, information contained in the discrepancies between these evolutionary histories can give us an insight into these ancestral genomic events which would provide efficient instruments in a range of fields, such as establishing orthology/paralogy relationships between gene families, functional gene annotations, reconstruction of ancestral genes and genomes and their dating, construction of phylogenies based on whole genome data, event-based reconstruction of coevolution and accurate reconstruction of gene and species trees (Rusin *et al.*, 2014).

In this thesis we are going to focus specifically on phylogenetic incongruence as a result of lateral gene transfers. We aim to reconstruct, not only the gene trees given a known species tree, but also the gene transfer history relating them.

A *Lateral Gene Transfer* (LGT) occurs when organisms from distinct species exchange genomic material directly, thereby breaking the usual pattern of inheritance-by-descent. It is an example of a biological mechanism by which the evolutionary tree for a particular gene can differ from the overall pattern of species evolution.

The presence of genes that have undergone LGT in a data set for species tree inference usually causes severe problems due to conflicting signals for the tree topology. Simulation studies (Beiko *et al.*, 2008) have shown the impact that LGT events can have on estimated phylogenetic trees. The authors used a sophisticated simulation procedure to evolve populations of genomes under various tree-like models of DNA substitution, and non-tree-like LGT events. Phylogenetic trees were then estimated using standard methods from the simulated genomic data, and compared to the underlying imposed evolutionary history. Overall, LGT had the effect of drastically decreasing statistical support for most relationships in the recovered tree.

Beyond tree construction, the correct identification of genes that have undergone LGT is an important biological problem, since it sheds light on the molecular pathways in which the genes play a role. Moreover, in prokaryotes, LGT is recognised as a major force allowing "evolution by acquisition" whereby new capabilities, including genes related to virulence, may be acquired quickly. Understanding and quantifying LGT in pathogens therefore has implications for the study of pathogen related disease in humans.

A principled statistical model for detecting LGT events in gene trees would involve a combination of the following hierarchical levels. At the top of the hierarchy, we would consider a model for the underlying species tree. A model of LGT, possibly incorporating prior biological information about the relative probabilities of different transfer events, would then be used to relate the species tree to different related gene trees. A model of sequence evolution over the individual gene trees would then relate the tree to the observed genetic sequence data. Fitting such a model would involve combined simultaneous inference of gene trees and their relationship to an underlying species tree.

Suchard (2005) developed a Bayesian approach to joint estimation of gene trees and an underlying species tree in the presence of LGT. This thesis draws heavily on Suchard's work. It consists of a hierarchical model, with different levels representing the unrooted species tree, gene trees, and gene sequence alignments. A random walk on the space of tree topologies was used to model the LGT process and relate gene trees to the species tree. It is assumed that the *subtree prune and regraft (SPR)* operator mirrors the observed effect that LGT has on inferred trees and that one step on the random walk represents one SPR on the current tree. Hierarchical phylogenetic models are used to reconstruct each gene tree from its corresponding gene sequence alignment, enabling the pooling of information across genes.

The overall aim of the project described in this thesis is to develop statistical methods which are capable of identifying, modelling and analysing the processes that give rise to variation caused by LGT in the inferred gene trees. To achieve this we propose a more biologically realistic approach to the Bayesian hierarchical model proposed by Suchard. Our model assumes the species tree as a fixed rooted tree and takes into account possible time orderings of divergence events in trees, without explicitly modelling divergence times. The time ordering becomes a natural constraint to LGT events by allowing them to happen only between species that are contemporary. As a result, an extended version of the SPR operator (xSPR), which respects the time ordering, is developed as a more adequate way of describing the effect of an LGT on a phylogenetic tree. Since we are working under a Bayesian framework, and given the strong history of its use in phylogenetics since the mid- to late-1990s, we use Markov Chain Monte Carlo (MCMC) methodologies for parameter inference and developed a novel proposal mechanism in order to jointly propose the LGT distance, gene transfer history and gene trees themselves. A proposal mechanism for ordering a phylogenetic tree is also introduced. Hierarchical phylogenetic models are used in the reconstruction of each gene tree from its corresponding gene sequence alignment, as in Suchard's model, but we further assume that variation in the rate of evolution between different sites is not constant but Gamma distributed. A novel graphical representation, LGT Biplot, is introduced as an adequate and easily understandable way to present the gene transfer history.

The thesis is organised as follows. Chapter 2 provides an overview about modelling molecular evolution, introducing models of nucleotide substitution and common approaches to gene tree and species tree inference, with special emphasis on Maximum Likelihood and Bayesian inference methods. Chapter 3 discusses phylogenetic incongruence derived by LGT events, with Subsection 3.2.1 and Section 3.3 describing the model proposed by Suchard (2005). In Chapter 4 our Bayesian hierarchical phylogenetic model is introduced in detail, as well as all the components related to performing Bayesian inference such as proposals and the MCMC algorithm. Chapters 5 and 6 shows an application of our model to simulated and real data. Finally, in Chapter 7 we present closing remarks and discuss open problems and possible extensions of this work.

There are several novel aspects to the research presented in this thesis. Unlike Suchard we assume that the species tree is a fixed rooted tree and that divergence events happened according to a time ordering. Under these assumptions, we introduce the xSPR operator which is a more adequate representative of an LGT in an ordered rooted tree, when assuming gene transfers only between contemporary species. MCMC proposal mechanisms are constructed for ordering a rooted phylogenetic tree and to jointly propose LGT distance, LGT history and gene trees. In relation to the substitution model we assume gamma rate heterogeneity, a fundamental improvement on Suchard's model. We also introduce a novel graphical representation of gene transfer history, the LGT Biplot.

2

Modelling Molecular Evolution

Reconstructing evolutionary relationships between species and investigating forces and mechanisms of the evolutionary process are the two major aims of studying evolution at the molecular level. Because molecular data is readily available and contains much more information that any other type of data available, it has become the most common type of data used for phylogeny reconstruction provoking a phenomenal growth in both areas of research for the last decades, motivated also by the improved computer hardware and software and development of sophisticated statistical methods (Yang, 2006). Initially, the available datasets consisted of multiple sequences alignments from only one specific gene and commonly used tree reconstruction methods would produce a single estimate of the tree. This tree was then treated as the estimated species tree, ignoring the variability of gene trees for genes evolving from the same fixed pattern of speciation, or species tree. Throughout this thesis we will assume that the phylogeny for a set of gene sequences from the species is called the *gene tree* while the phylogeny representing the relationships, specifically the strict branching pattern of divergence among a group of species is the *species tree.*

Although the variability and incongruence in estimated evolutionary trees (phylogenies) has attracted considerable research interest recently and promising approaches have been proposed, gene tree estimation continues to play a central role in the current species tree reconstruction techniques, so it is worthwhile to begin by understanding not only the methods behind gene tree inference but also to introduce some essential concepts on genetic variation and evolution.

Variation and Evolution

Different species have different genomes and within each species, DNA sequences differ between individuals. The exact *genotype* of an individual, i.e. the DNA sequence corresponding to a gene, determines its observable characteristics, also known as *phenotype*. Variation in genotype will affect the way proteins are constructed, in terms of their structure and timing of their production.

As a result of sexual reproduction (e.g. in mammals), organisms inherit two copies of each chromosome, one from each parent, and thereby obtain the characteristics of their parents. This means they possess two copies of each gene leading to different combinations of genes in their offspring. Both nucleotide sequences will often differ and might even produce proteins that will result in different phenotypes. Although simpler forms of organisms (e.g. bacteria) might originate offspring genetically very similar to their parents, their sequences will be always subject to the fact that errors (mutations) occur during genome replication. In its simplest form, a mutation will be a single nucleotide change which might radically change the structure of the coded protein affecting its function. Mutations occur completely at random and usually at a relatively low rate. Many mutations will be disadvantageous in terms of the survival of the offspring. Others will have little or no effect on the organism and these are referred to as neutral. Advantageous mutations will be the ones that increase the fitness of the organism enabling it to survive longer and reproduce. DNA replication and random mutation are two fundamental concepts of evolution. Together they lead to a slow accumulation of mutations within a population which results in evolution of the corresponding characteristics.

The concept of natural selection has a direct connection with evolution and describes the process in which organisms, within a population, with certain genetic traits are more likely to reproduce than others, increasing the likelihood of passing their DNA onto the next generation. It acts as a filter between the genotypes of subsequent generations and it operates at a probabilistic level in the sense that it affects the relative probabilities of genotypes. Without the process of mutation, natural selection would eventually result in some sort of equilibrium in the form of an unchanging population, but mutation usually ensures that the offspring genotype differs from that of its parents.

An initial advantageous or neutral mutation will occur in only one individual and if genetic drift does not prevent it from spreading through a population, this mutation might be transmitted to its offspring in reproduction. The offspring will carry and pass the mutation to its offspring and the process carries on until it becomes present in a large proportion of the population, in which case we say it is *fixed*. When the mutation corresponds to a change in one nucleotide it is called *substitution*. An accumulation of substitutions and/or other forms of mutations might lead to a *speciation event*. A *species* is often defined as a collection of organisms that are capable of interbreeding and producing fertile offspring. When mutations accumulate in sub-populations of a fixed species, those sub-populations might diverge and become unable to mate, originating a new species.

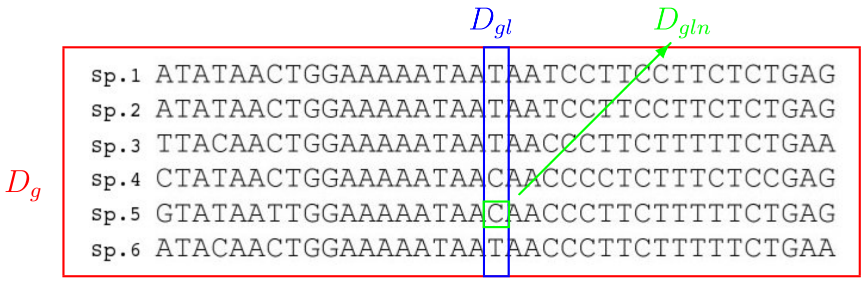
Inferring phylogenetic gene trees

Phylogenetic gene tree inference consists broadly of two steps: a) obtaining and aligning molecular sequences and b) inferring gene trees for the aligned sequences. In the first step, multiple sequence alignment has become an indispensable tool in modern molecular biology research. The DNA sequences to be aligned often contain open reading frames (ORFs) which are a section of a sequenced piece of DNA that begins with an initiation (methionine ATG) codon and ends with a nonsense/stop codon. These sequences code for proteins and therefore a coding sequence can be considered either at the nucleotide or amino acid level. Due to the redundancy of genetic codes, different codons, i.e. a sequence of three DNA or RNA nucleotides that corresponds with a specific amino acid or stop signal during protein synthesis, encode the same amino acids. Therefore, a nucleotide sequence is less conserved but more informative than its amino acid translation. For the purpose of this thesis, we will analyse uniquely DNA sequences. Probabilistic sequence alignment models have been shown to provide an effective framework for building accurate sequence alignment tools. Numerous tools exist to align DNA sequences, among which are CLUSTAL (Higgins *et al.*, 1992), T-COFFEE (Notredame *et al.*, 2000), DIALIGN (Morgenstern *et al.*, 2002), MUSCLE (Edgar, 2004), MAFFT (Katoh *et al.*, 2005), FSA (Bradley *et al.*, 2009), PRANK (Löytynoja & Goldman, 2008), and the more recently proposed MACSE (Ranwez *et al.*, 2011). The initial alignment can strongly impact conclusions and biological interpretations (Wong *et al.*, 2008).

Despite its importance, multiple sequence alignment is not in the scope of this thesis. Our main concern lies in achieving accurate gene tree inference for the species in study, given a fixed multiple sequence alignment. In order to set a basis for the nomenclature that will be used throughout this thesis let us first formally describe an example for a single gene phylogenetic analysis.

Suppose we intend to analyse the evolutionary relationships between a set of N

species using a multiple sequence alignment of N molecular sequences for gene g, whose length of aligned sites (number of columns) we denote L. Site data $D_{gl} = (D_{gl1}, \ldots, D_{glN})$ contains one nucleotide from each taxon, such that $D_{gln} \in \boldsymbol{\mathcal{A}}$, for $\boldsymbol{\mathcal{A}} = (A, C, G, T)$, $l = 1, \ldots, L$ and $n = 1, \ldots, N$ (see Figure 2.1).



```
sp.1 ATATAACTGGAAAAATAATAATCCTTCCTTCTCTGAG
sp.2 ATATAACTGGAAAAATAATAATCCTTCCTTCTCTGAG
sp.3 TTACAACTGGAAAAATAATAACCCTTCTTTTTCTGAA
sp.4 CTATAACTGGAAAAATAACAACCCCTCTTTCTCCGAG
sp.5 GTATAATTGGAAAAATAACAACCCTTCTTTTTCTGAG
sp.6 ATACAACTGGAAAAATAATAACCCTTCTTTTTCTGAA
```

Figure 2.1: **Example of a multiple sequence alignment D_g for gene g, with $N = 6$ and $L = 37$.** The D_{gl} column (blue) represents the set of nucleotides (one for each species) aligned at position l, while D_{gln} (green) corresponds specifically to the nucleotide in position l belonging to species n.

Phylogenetic trees can describe the genealogical relationships among species, genes, populations or even individuals. Here we consider that the leaves represent present-day species, internal vertices represent extinct ancestral species with no sequence data available, and the branching pattern shows how species have diverged. The most recent ancestor of all species is the *root* of the tree. The existence of a distinguished point on a tree enables us to define a sense of time direction, that is, time flows from the root to the leaves (Song, 2003). The branching pattern of a tree is called *topology* and the lengths of its branches may represent the amount of sequence divergence or the time period covered by the branch (Yang, 2006). The number of branches connected to a vertex is called the *degree* of the vertex.

Mathematically, a *rooted phylogenetic tree* is a directed acyclic graph with a unique vertex corresponding to the most recent common ancestor of all the entities at the leaves of the tree. Formally, an N-taxon rooted phylogenetic tree $T = (\tau, \ell)$ has N labelled degree-1 vertices, $N - 2$ unlabelled degree-3 vertices, and a distinguished vertex of degree-2 called the *root.* The tree has $2N-2$ edges, each of which has a non-negative length ℓ_j, for $j = 1, ..., 2N-2$, where $\ell = (\ell_1, ..., \ell_{2N-2})$ is the vector of edge lengths. In a tree T, time flows from the root to the leaves, with edges oriented to reflect this. A directed *path* from vertex v_0 to vertex v_k is an alternating sequence of vertices $v_0, \ldots, v_k$ and edges $e_1, \ldots, e_k$, such that e_i joins v_{i-1} and v_i, and all e_is and v_is are distinct. We say that a vertex v is a *descendant* of vertex u if there exists a path from u to v which goes strictly forward in time; u is called an *ancestor* of v. The set $\{v_1, v_2, \ldots, v_{N-2}\}$ of degree-3 vertices is a partially ordered set since the ancestor-descendant relations between vertices defines the ordering of some vertices. If we cut an edge on a tree, two subtrees will be originated, dividing the species into two mutually exclusive sets. Thus, each edge on a tree defines a *bipartition* or *split* of the species.

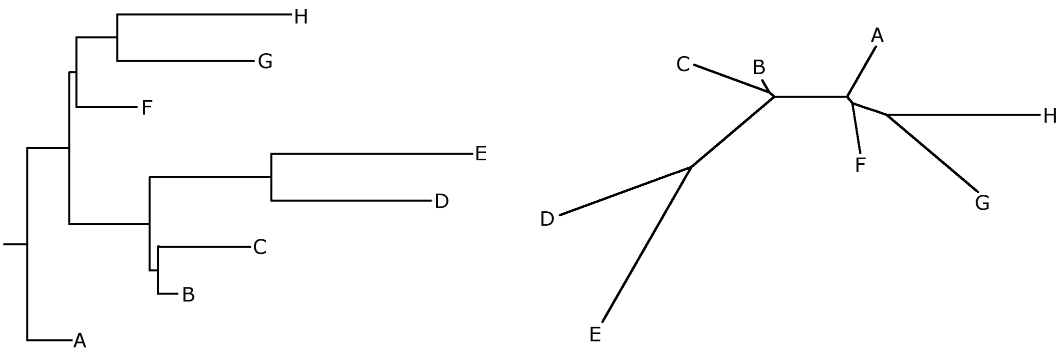
Trees are often represented using the Newick format's parenthesis notation. Sister taxa are grouped into one clade (which represents the set of species descended from a particular ancestral species) using parentheses, branch lengths are prefixed by colons and a semicolon marks the end of the tree.

As an example, the phylogenetic tree in Figure 2.2(a) is a graphical representation of the following Newick string:

(A:0.3185656,(F:0.1462084,(G:0.0048672,H:0.0306487):0.1396449):0.0077896,
((B:0.0212217,C:0.2387582):0.0826425(D:0.1449661,E:0.0199936):0.2650590):
0.0320244);

It represents the evolutionary relationships of 8 hypothetical species as a *phylogram* with branches drawn proportionally to their lengths and measured by the expected number of nucleotide substitutions per site.

Often insufficient information exists to determine the root and the tree is left unrooted (Figure 2.2(b)) still providing a notion of the evolutionary relationships between organisms (Hickey *et al.*, 2008).



(a) Rooted phylogenetic tree.

(b) Unrooted phylogenetic tree.

Figure 2.2: Rooted and unrooted phylograms of the example Newick string.

The tree represented in Figure 2.2(a) is a *binary* or *bifurcating* tree. If the root vertex had a degree greater than 2 or a nonroot vertex had a degree greater than 3, then that vertex would represent a *polytomy* and we would say we were in the presence of a *multifurcating* tree. For the purpose of this thesis we will assume only bifurcating trees.

The concept of *tree space*, which is the set of all possible N taxa trees is also key in phylogenetics. The total number of unrooted bifurcating trees for N taxa (Z_N) is easily seen to be

$$Z_N = 1 \times 3 \times 5 \times 7 \times \ldots \times (2N - 5), \tag{2.1}$$

by an inductive argument (Cavalli-Sforza & Edwards, 1967).

Considering that each unrooted tree has $(2N-3)$ branches, with the possibility of placing the root on any of those branches, there will be $(2N-3)$ rooted trees for each unrooted tree. Therefore, the number of rooted trees for N species is $Z_N \times (2N-3)$.

This number increases exponentially with the number of species (Yang, 2006).

2.1 Markov Models Of Nucleotide Substitution

A variety of evolutionary models to explain the mechanism of nucleotide or amino acid substitutions have been proposed. Calculation of the distance between two sequences is perhaps the simplest analysis, being the first step in distance-matrix methods of phylogeny reconstruction. The distance between two sequences can be defined as the average number of nucleotide changes per site, meaning that when the evolutionary rate is constant over time, the distance will increase approximately linearly with the time of divergence. The estimation of this distance can be obtained through the use of a probabilistic model, commonly a continuous-time Markov chain, to describe substitutions. Assuming that the nucleotide sites in the sequence evolve independently of each other, substitutions at any site are described by a Markov chain with the four nucleotides as the states of the chain. As a result of the Markovian property, the probability in which the chain jumps into other nucleotide states depends on the current state, but not on how the current state is reached (Yang, 2006).

Following the formal description of the substitution process in Yang (2006), let $X(t)$ be the state of the chain at time t corresponding to one of the four nucleotides A, C, G and T. We further assume that different sites in a DNA sequence evolve independently and the same Markov-chain model describes the nucleotide substitutions at any site. The chain is characterised by the rate matrix $\boldsymbol{Q} = (q_{ij})$, where q_{ij} is the instantaneous rate of change from i to j such that $Pr\{X(t + \mathit{\Delta} t) = j|X(t) = i\} \simeq q_{ij}\mathit{\Delta} t$, for any $j \neq i$ and small $\mathit{\Delta} t$. As we assume that q_{ij} does not depend on time this is a *time-homogeneous* process with q_{ii} specified in such a way that each row of $\boldsymbol{Q}$ sums to zero, $q_{ii} = -\sum_{j\neq i} q_{ij}$. The $\boldsymbol{Q}$ matrix determines the *transition-probability matrix* over any time $t > 0 : \boldsymbol{P}(t) = (p_{ij}(t))$, where $p_{ij}(t) = Pr\{X(t) = j|X(0) = i\}$. Standard theory of Markov processes (Yang, 2006) shows that

$$\boldsymbol{P}(t) = \exp(\boldsymbol{Q}t). \tag{2.2}$$

Assuming that the Markov chain $X(t)$ has the initial distribution

$$\pi^{(0)} = (\pi_A^{(0)}, \pi_C^{(0)}, \pi_G^{(0)}, \pi_T^{(0)}), \tag{2.3}$$

the distribution at time t is

$$\pi^{(t)} = (\pi_A^{(t)}, \pi_C^{(t)}, \pi_G^{(t)}, \pi_T^{(t)}) = \pi^{(0)}\boldsymbol{P}(t). \tag{2.4}$$

A stationary distribution is a distribution π for $X(t)$ such that $\pi(t) = \pi \implies \pi(s) = \pi, \forall\ s \geq t$. This Markov chain has a unique stationary distribution $\pi\boldsymbol{P}(t) = \pi \Leftrightarrow \pi\boldsymbol{Q} = 0$, which is also the *limiting distribution* when time $t \to \infty$. The chain is also assumed to be *irreducible* allowing any state to change into any other state in finite time.

By placing further constraints on substitution rates between nucleotides, different models of nucleotide substitution are created. These substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution. The simplest models, JC69 (Jukes & Cantor, 1969) and K80 (Kimura, 1980), have symmetrical substitution rates, $q_{ij} = q_{ji}$ for all i and j, with $\pi_i = 1/4$. The first assumes that every nucleotide has the same rate of changing into any other nucleotide, while the latter assigns different rates for *transitions* (substitutions between pyrimidines ($T \leftrightarrow C$) or purines ($A \leftrightarrow G$)) and *transversions* (substitutions between a pyrimidine and a purine ($A, G \leftrightarrow T, C$)). The Tamura and Nei (1993) TN93 model accommodates unequal base composition $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$ with different rates for transversions, transitions between pyrimidines and transitions between purines. A special case of TN93, which we will use throughout this thesis, is known as HKY85 and was proposed by Hasegawa *et al.* (1985). It is parameterised also by unequal base compositions and, as in the K80 model, it assumes different rates for transitions (γ) and transversions (δ), which can be translated by the transition:transversion rate ratio $\rho = \gamma/\delta$.

Given this parameterisation, the HKY85 model rate matrix is given by

$$\boldsymbol{Q} = \delta \begin{pmatrix} - & \pi_C & \rho\pi_G & \rho\pi_T \\ \pi_A & - & \pi_G & \rho\pi_T \\ \rho\pi_A & \pi_C & - & \pi_T \\ \pi_A & \rho\pi_C & \pi_G & - \end{pmatrix}, \tag{2.5}$$

where matrix indices reflect the nucleotides ordered as A,C,G,T. The elements denoted by a dash take values to ensure the row sum is zero. A detailed description of the results for the TN93 model, which also applies to HKY85, as well as information on more complex models, such as the General Time Reversible (GTR) model, which allows for different nucleotide frequencies and 6 different substitution rates, can be found in Yang (2006).

The *molecular evolutionary clock* hypothesis states that the rate of DNA or protein sequence evolution is constant over time or among evolutionary lineages (Zuckerkandl & Pauling, 1965). This hypothesis had a tremendous impact in the field of molecular evolution by allowing inference on divergence times among species. But controversy arose around its reliability and implications for the mechanism of molecular evolution; see Yang (2006) for a brief review of the debate.

In this thesis we will not assume that the evolutionary rate of DNA sequences is constant over time. Instead we assume no clock-like restrictions on branch lengths. The non-clock model is the standard model used in phylogenetic inference. It can be considered as a branch-breaking model which allows the evolutionary rate to be different for each branch in the tree. As a result, the non-clock tree model has a large number of free parameters, one for each branch in the tree. If there are N leaves in the tree, there are $2N - 3$ branches and hence branch length parameters in the tree model. This means that, for a typical phylogenetic problem, the largest number of free parameters comes from the branch lengths.

It has long been recognized that it is unreasonable to assume a constant overall rate of evolution across nucleotide sites in a sequence. Failure to account for variation in rates across sites would result in underestimating the sequence distance (Yang, 2006) and, under some conditions, lead to phylogenetic artifacts such as long-branch attraction, as has been shown in simulation studies and in analyses of real data (Susko *et al.*, 2003). Although the models mentioned above assume rate homogeneity, heterogeneity can be accommodated by assuming that the specific rate r for each site is a random variable drawn from a Gamma distribution with density function

$$g(r|\alpha, \beta) = \frac{\beta^\alpha}{\mathit{\Gamma}(\alpha)} r^{\alpha-1} \mathrm{e}^{-\beta r}, \qquad r > 0, \tag{2.6}$$

where $\alpha > 0$ and $\beta > 0$ are the shape and scale parameters. It is a common procedure to set $\beta = \alpha$, so that the mean $\beta/\alpha = 1$ with variance $\alpha/\beta^2 = 1/\alpha$. In this way, for $\alpha \leq 1$ the gamma distribution will have a skewed L-shape which means most sites will have very low rates of evolution while an $\alpha > 1$ describes a bell-shaped distribution with most sites having an intermediate rate around 1 (see Figure 2.3). Assuming variable rates among sites, the sequence distance is defined as the expected number of substitutions per site, averaged over all sites (Yang, 2006).

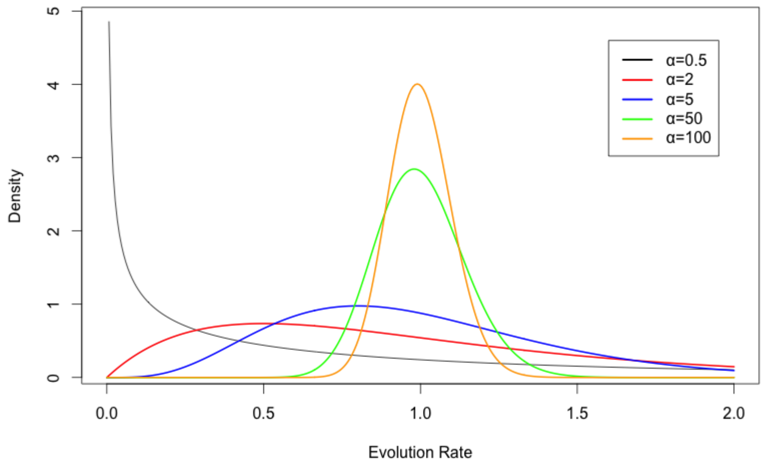


Figure 2.3: **Gamma distribution probability density function for variable rates among sites.** The scale parameter β is equal to the shape parameter α so that the mean of the distribution is 1.

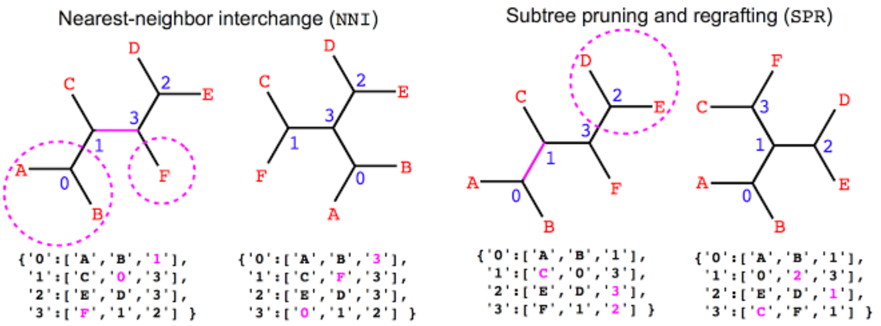
Therefore, $P_{ij} = (\exp\{\ell_k \boldsymbol{Q}\})_{ij}$, where P_{ij} is the probability of nucleotide i being replaced by nucleotide j and $\ell_k = t \times r$. Conventionally, $\boldsymbol{Q}$ is multiplied by a scale factor so that the average rate is 1 and time t is measured by the expected number of substitutions per site.

2.2 Phylogeny Reconstruction

Methods for inferring gene trees from sequence data are numerous and have become extraordinarily sophisticated in recent years. The most common traditional approaches are *distance-based* methods such as the Neighbor-Joining algorithm (Saitou & Nei, 1987) and UPGMA - unweighted pair-group method using arithmetic averages (Sokal & Michener, 1958; Murtagh, 1984), in which distances are calculated from pairwise comparison of sequences, and a clustering algorithm is usually used to convert the distance matrix into a phylogenetic tree. On the other hand, *character-based* methods attempt to fit the nucleotides or amino acids observed in all species at every sequence site to a tree. These methods include Maximum Parsimony (Fitch, 1971; Hartigan, 1973), Maximum Likelihood (ML) (Felsenstein, 1973a,b) and more recently, Bayesian approaches which have provided a method for simultaneously obtaining trees and measurements of uncertainty for every parameter (Mau & Newton, 1997; Yang & Rannala, 1997; Holder & Lewis, 2003). The Markov-process models of nucleotide substitution used in distance calculation form the basis of likelihood and Bayesian analysis of multiple sequences on a phylogeny.

Methods for tree reconstruction are usually classified into two types: *algorithmic* (cluster methods) or *optimality based* (search methods). Neighbor-Joining and UPGMA are included in the first group as they use cluster algorithms in order to obtain a single tree as the estimate of the true tree. In optimality-based methods, the tree with the optimal score, according to an optimality criterion, is the estimate of the true tree (Yang, 2006). The maximum parsimony method assumes that the *most parsimonious tree* is the tree that requires the minimum number of evolutionary steps (character changes) to explain the observed pattern. In the ML method, the *maximum likelihood tree* is the tree with the highest likelihood value, under a probabilistic model of nucleotide substitution, when fitted to the data. Bayesian inference generates a posterior distribution for the model parameters, composed of a phylogenetic tree and a model of evolution, based on the prior for that parameter and the likelihood of the data, given a multiple sequence alignment. The fact that Bayesian inference infers a distribution for a parameter, accounting also for uncertainty, makes its classification as an optimality-based method controversial, although it might be considered as such when the aim is to obtain the tree with the maximum posterior probability, known as the *MAP tree*. Both ML and Bayesian approaches are model-based, using substitution models to calculate the likelihood function.

When evaluating trees according to an optimality criterion, theoretically, the score for every tree can be calculated and the tree with the best score can be identified. Although this *exhaustive search* can lead to the best tree, it is computationally unfeasible for larger datasets (e.g. for $N = 20$, the number of possible unrooted trees is 2.22×10^{20}). Therefore some heuristic algorithms are needed to search in tree space. There are two categories of heuristic search algorithms: *hierarchical clustering algorithms* (see Yang (2006) for detailed information) and *tree-rearrangement* or *branch-swapping* algorithms which propose new trees from an initial one, through local perturbations to the current tree and generating a collection of neighbour trees. The decision of moving or not to a new tree will depend on the inference methods in use.


Nearest-neighbor interchange (NNI)
Subtree pruning and regrafting (SPR)
A
B
C
D
E
F
0
1
2
3

```
{'0':['A','B','1'],
 '1':['C','0','3'],
 '2':['E','D','3'],
 '3':['F','1','2'] }
```

```
{'0':['A','B','3'],
 '1':['C','F','3'],
 '2':['E','D','3'],
 '3':['0','1','2'] }
```

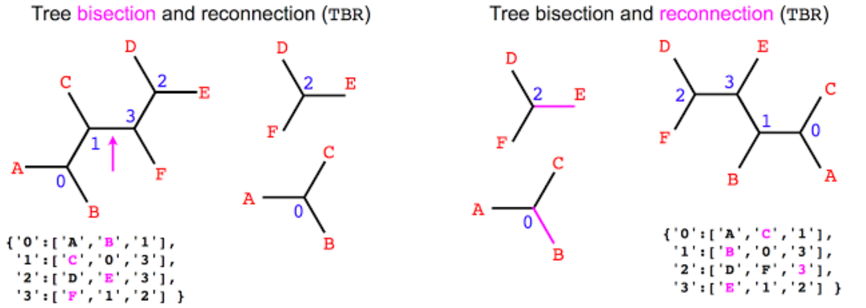
(a) NNI operator.

```
{'0':['A','B','1'],
 '1':['C','0','3'],
 '2':['E','D','3'],
 '3':['F','1','2'] }
```

```
{'0':['A','B','1'],
 '1':['0','2','3'],
 '2':['E','D','1'],
 '3':['C','F','1'] }
```

(b) SPR operator.

Several branch-swapping algorithms have been proposed. The *nearest-neighbour interchange (NNI)* uses the fact that each internal branch defines a relationship among four subtrees (Figure 2.4(a)). It proposes a move by swapping a subtree on one side of an internal branch with a subtree on the other side. Another mechanism is the *subtree prune and regraft (SPR)* operation which is defined as cutting any branch and thereby pruning a subtree, and then regrafting the subtree by the same cut edge to a new vertex obtained by subdividing another edge in the tree (Figure 2.4(b)). A


Tree bisection and reconnection (TBR)
D
C
2
E
3
1
A
0
F
B
D
2
E
F
C
A
0
B
Tree bisection and reconnection (TBR)
D
2
E
F
C
A
0
B
D
E
2
3
C
F
1
0
B
A

```
{'0':['A','B','1'],
 '1':['C','0','3'],
 '2':['D','E','3'],
 '3':['F','1','2'] }
```

```
{'0':['A','C','1'],
 '1':['B','0','3'],
 '2':['D','F','3'],
 '3':['E','1','2'] }
```

(c) TBR operator - bisection.

(d) TBR operator - reconnection.

Figure 2.4: Branch-swapping mechanisms (Elliott, 2014).

greater perturbation is obtained through the *tree bisection and reconnection (TBR)* where after cutting the tree in two subtrees by pruning one internal branch (Figure 2.4(c)), a new tree is formed by choosing and rejoining two branches, one in each tree (Figure 2.4(d)). NNI generates fewer neighbours than SPR, which generates fewer neighbours than TBR (Allen & Steel, 2001).

A characteristic of objective functions on tree space is the existence of *local peaks* or *tree islands*. Multiple local peaks are known to exist in the tree space during heuristic tree search under maximum parsimony and ML (see Yang (2006) for a practical example). Steel (1994) demonstrated that the maximum likelihood point for a phylogenetic tree is not necessarily unique, a phenomenon that had been encountered by Cavalli-Sforza & Edwards (1967) in their early attempts to apply maximum likelihood in phylogenetic tree reconstruction (Rogers & Swofford, 1999). Due to this, and considering that branch swapping is a hill climbing algorithm, it is able to find a local optimum, but such a solution cannot be improved by considering a neighbouring tree. Branch swapping does not guarantee finding the global optimum out of all possible solutions (the tree space), and this becomes more serious for larger trees with more species (as the tree space is much larger) or longer sequences (with more sites, which tends to originate higher peaks and deeper valleys making it very difficult to traverse between peaks) (Yang, 2006). The same problem can be expected for *stochastic tree search*, although some stochastic search algorithms attempt to overcome the problem of local peaks by allowing downhill moves, as with *simulated annealing* (Metropolis *et al.*, 1953; Kirkpatrick *et al.*, 1983). In this optimisation algorithm, the objective function is modified (heated) to have a flattened surface in the early stage of the search facilitating moves between peaks by allowing a higher acceptance of downhill moves. As the simulation proceeds the 'temperature' is gradually reduced and at the final stage of the algorithm, only uphill moves of the algorithm are accepted (Yang, 2006). See Barker (2004) for an example of the use of simulated annealing for phylogenetic inference. The *genetic algorithm* is another stochastic tree-search algorithm mainly used for ML tree search which uses operations similar to mutation and recombination to generate new trees from the current ones. Each tree will be kept in every generation depending on a 'fitness'-related optimality criterion. An example of such an algorithm can be found in Lewis (1998).

The stochastic tree-search algorithm that we will give particular attention to is the Markov Chain Monte Carlo (MCMC) algorithm. This algorithm has the advantage of being a statistical approach which assigns a posterior probability for each tree in the posterior distribution obtained when the chain reaches equilibrium. The Metropolis-Coupled Markov Chain Monte Carlo (MCMCMC) algorithm, introduced by Geyer (1991) for multimodal distributions is a parallel-chain extension of this algorithm which uses some of the simulated annealing algorithm principles allowing multiple peaks in the landscape of trees to be more readily explored. Subsection 2.2.2 contains a more detailed view of both algorithms. For the purpose of this thesis we will next overview the theoretical concepts behind inferring molecular evolution using ML and Bayesian approaches.

2.2.1 Maximum Likelihood

In 1981, Felsenstein provided a likelihood function for a gene sequence alignment given a gene tree, in terms of a Markov substitution model. Then, the estimate of the gene tree is obtained by maximizing the function with respect to the gene tree. He assumed that sites within a gene are independent and identically distributed and also that evolution in one lineage is independent of other lineages. Thus, the probability of observing D_{gl} is given by a multinomial distribution over the 4^N possible outcomes. The multinomial probabilities are functions of an unknown rooted binary tree topology τ, branch lengths $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_B)^T$, for $B = 2N - 2$, and a model to describe the mutation of nucleotides along the branch lengths.

Let us consider the HKY85 model referred to earlier with the substitution-rate matrix presented in Equation 2.5. Since only the product $\ell_b \times \boldsymbol{Q}$ enters into the model likelihood, we fix

$$\delta = \frac{1}{2[\rho(\pi_A\pi_G + \pi_C\pi_T) + (\pi_A + \pi_G)(\pi_C + \pi_T)]}, \tag{2.7}$$

and this constraint will enforce that

$$\sum_{m \in (A,G,C,T)} = \pi_m \boldsymbol{Q}_{m,m} = -1, \tag{2.8}$$

so that each branch length is the expected number of nucleotide substitutions per site between the two vertices the branch connects.

Considering the set of parameters $\boldsymbol{\Psi} = (\boldsymbol{Q}, \boldsymbol{\ell}, \tau)$, the likelihood with respect to alignment D, which has a total length of L independent and identically distributed sites, is

$$p(D|\boldsymbol{\Psi}) = \prod_{l=1}^{L} p(D_l|\boldsymbol{\Psi}). \tag{2.9}$$

Equivalently the log-likelihood is a sum over the sites in the sequence

$$\log\ p(D|\boldsymbol{\Psi}) = \sum_{l=1}^{L} \log\ p(D_l|\boldsymbol{\Psi}). \tag{2.10}$$

Variation in the rate of evolution between different sites can be accommodated in the likelihood function using the *discrete gamma method* of Yang (1993, 1994). A discrete approximation is usually used for continuous rate distribution models such as the gamma because of the computational difficulties in evaluating likelihoods. Thus, the distribution is partitioned into c "rate categories" of equal probability, each with a rate constant r_y for $y = 1, ..., c$. The probability $p(D_l|\boldsymbol{\Psi})$ is approximated as

$$p(D_l|\boldsymbol{\Psi}) \simeq \frac{1}{c}\sum_{y=1}^{c} p(D_l|\mathbf{Q}, \tau, r_y\boldsymbol{\ell}). \tag{2.11}$$

The probability of each individual column, $p(D_l|\mathbf{Q}, \tau, r_y\boldsymbol{\ell})$ will be the sum of the probabilities over all possible nucleotide combinations for the extinct ancestors, represented by the interior vertices of the tree. This calculation is computationally expensive as there are 4^{n-1} possible combinations for $n-1$ interior vertices. Felsenstein's "pruning" algorithm (Felsenstein, 1973a, 1981), is the most common approach for such a calculation.

The root of the tree can be placed anywhere on the tree without affecting the likelihood. Therefore, we are estimating not a single rooted tree but an equivalence class of rooted trees, namely all those compatible with a given unrooted tree, which is what we are in effect estimating (Felsenstein, 1981).

Once we know how to calculate the likelihood, the estimation becomes a standard mathematical problem: maximizing the likelihood function over the entire parameter space. The parameters here include not only gene trees but also the parameters in the substitution models and any parameters used to model the correlation among sites. A detailed explanation of numerical algorithms for maximum likelihood estimation can be found in Yang (2006).

Maximum likelihood estimates are generally consistent in that the estimates converge to the true gene trees as the length of the gene sequences goes to infinity, if the underlying model is correct. Statistical tests are often employed to find the best model for the data. The likelihood ratio test is one of the most commonly used techniques to select the model, from among a hierarchy of models, that most appropriately fits the data. However, choosing the model with the highest likelihood value may lead to one that is unnecessarily complex. In addition, it is inappropriate to use the likelihood ratio test to select the topology of the gene tree because the likelihood ratio test requires the models to be nested. This has led many investigators to consider model selection criteria such as the Akaike information criterion (AIC) (Akaike, 1974), which does not require nested models. The AIC score is calculated for each model, defined as

$$AIC = -2 \log L + 2p \tag{2.12}$$

where $\log L$ is the optimum log-likelihood that measures the goodness of fit of the model and p the number of parameters. Another commonly used model selection approach, cross validation, is based on minimizing the prediction error. However, cross validation involves intensive computation, which dramatically limits its application to tree building projects.

2.2.2 Bayesian Inference

Bayesian approaches to phylogenetic analysis are relatively new, being introduced in the late 1990's (Rannala & Yang, 1996; Yang & Rannala, 1997; Mau & Newton, 1997; Li *et al.*, 2000). Initially, a constant rate of evolution was assumed (the molecular clock) as well as a flat prior on rooted tree topologies. But more efficient algorithms have been implemented since then in which the clock constraint is relaxed, enabling phylogenetic inference under more realistic evolutionary models (Yang, 2006).

The field of Bayesian statistics is closely allied with ML. The Bayesian method is different from the likelihood method in that it treats parameters as random variables and assumes prior distributions on them whereas parameters are unknown fixed constants in the likelihood paradigm. The notion that parameters have a probability distribution is key in Bayesian statistics being *per se* a measure of uncertainty for the parameters. In phylogenetic inference, the posterior probability can be interpreted as the probability that the tree is correct given the multiple sequence alignment, and assuming the multiple sequence alignment was generated under the model (i.e. no model mis-specification). Note that all ML analyses are also conditional on the model being correct. Let $\boldsymbol{\Psi}$ denote all the model parameters, including the unknown phylogenetic tree. Then the posterior distribution for the phylogenetic tree is obtained by combining the information in the data D, an aligned set of molecular sequences described by the likelihood function $p(D|\boldsymbol{\Psi})$, with that in the prior distribution, $p(\boldsymbol{\Psi})$.

Using Bayes Theorem, the posterior distribution is defined as

$$p(\boldsymbol{\Psi}|D) = \frac{p(\boldsymbol{\Psi})\, p(D|\boldsymbol{\Psi})}{p(D)}. \tag{2.13}$$

The likelihood is calculated under one of a number of standard nucleotide substitution models (referred to in Section 2.1) and the *marginal probability* of the data, $p(D)$, is a normalising constant that allows $p(\boldsymbol{\Psi}|D)$ to integrate to 1. Computation of $p(D)$ involves a summation over all trees topologies, and an integration over all possible combinations of branch length and substitution model parameter values which is analytically intractable. Fortunately, a number of numerical methods are available that allow the posterior probability of a tree to be approximated, the most commonly used being Markov Chain Monte Carlo (MCMC) which entails simulation from the posterior density $p(\boldsymbol{\Psi}|D)$.

Inferences about the history of the group of sequences are then based on the posterior probability of the phylogenetic model parameters. The mean, median or mode of the distribution can be used as equivalents to classical statistical "point estimates", and, for example, the *95% credible interval (CI)* can be constructed via the 2.5% and the 97.5% quantiles of the posterior density. Credible intervals can be *equal-tail* or *highest posterior density* depending on whether the posterior density is symmetric or skewed.

With the Bayesian approach, integration or marginalisation can be used to deal with nuisance parameters. Let us assume that $\boldsymbol{\Psi} = (\tau, \boldsymbol{\ell}, \boldsymbol{\theta})$ where τ is a topology such that $\tau \in \{\tau_1, \ldots, \tau_{t_n}\}$ with t_n as the total number of possible topologies for n species, $\boldsymbol{\ell}$ is the branch lengths on τ, and $\boldsymbol{\theta}$ corresponds to the substitution model parameters. The marginal posterior probability function of τ is

$$p(\tau|D) = \frac{\int_{\boldsymbol{\ell}} \int_{\boldsymbol{\theta}} p(D|\tau, \boldsymbol{\ell}, \boldsymbol{\theta}) p(\tau, \boldsymbol{\ell}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\ell} \mathrm{d}\boldsymbol{\theta}}{\sum_{j=1}^{t_n} \int_{\ell_j} \int_{\boldsymbol{\theta}} p(D|\tau_j, \ell_j, \boldsymbol{\theta}) p(\tau_j, \ell_j, \boldsymbol{\theta}) \mathrm{d}\ell_j \mathrm{d}\boldsymbol{\theta}} \tag{2.14}$$

In this case, τ is our parameter of interest and all other parameters are considered as nuisance parameters. The same reasoning can be applied to any of the other parameters.

Prior distribution

A prior probability distribution of an uncertain quantity ξ is the probability distribution that describes the uncertainty about ξ before taking into account any data D. There are three main ways to define a prior. A first approach is a model-based prior as, for example, the Yule process. The Yule process is a birth process with a constant birth rate. The number of births in time interval $(0, t)$, $Y(t)$, has a negative binomial distribution, which is the same as the distribution of the number of new species of a genus produced during $(0, t)$ in Yule's study of evolution. Another approach includes specifying the prior by assessing prior evidence concerning the parameter (e.g. using past observations of the parameters in similar circumstances) and using subjective beliefs of the researcher. *Vague* or *diffuse* priors are often used when little information is available. For example, in the absence of background data, a simple solution would be to assign equal probability to the possible trees, but it is always important to assess whether the posterior is sensitive to the prior. When the posterior is dominated by the information from the data, the choice of prior might be less important. When this is not the case, the effect should be assessed carefully.

The most common prior probability distributions used in Bayesian phylogenetics parameter estimation are the discrete and continuous Uniform, Exponential, Gamma, Normal and Dirichlet distributions. Specifically for the model we are fitting, Poisson and Geometric distributions are useful priors when modelling the number of times a certain event occurs. The main characteristics of these key probability distributions are reviewed in Appendix A.

An important concept in relation to priors is that of a *conjugate prior* which means that the prior and the posterior have the same distributional form and the information in the data is used to update the parameters in that distribution. Conjugate priors are convenient as the integrals are tractable analytically, but are not available for all distributions. As an example, if we assume that random variable X can be modelled by a normal distribution such that $X|\mu \sim N(\mu, \sigma^2)$, for known σ^2, the conjugate prior for μ is a Normal distribution $N(m, v)$ with density

$$p(\mu|m, v) \propto \frac{1}{\sqrt{v}} \exp\left\{ -\frac{1}{2v}(\mu - m)^2 \right\}, \qquad -\infty < \mu < \infty. \tag{2.15}$$

Since μ and X have a joint Normal distribution, the posterior distribution is

$$\mu|X = x \sim N\left(m + \frac{v}{\sigma^2 + v}(x - m), \left(\frac{1}{v} + \frac{1}{\sigma^2}\right)^{-1}\right). \tag{2.16}$$

On the other hand, the conjugate prior for σ^2, assuming fixed μ, is an Inverse Gamma $IG(a, b)$ with density

$$p(\sigma^2|a, b) \propto \frac{b^a}{\Gamma(a)} \sigma^{2(-a-1)} \exp\left(-\frac{b}{\sigma^2}\right), \qquad \sigma^2 > 0. \tag{2.17}$$

If the prior distribution involves unknown parameters, priors can be assigned to them, and in the same way, if these second-level priors contain other unknown parameters, these can have their own priors too. This is known as the *hierarchical Bayesian* approach. Usually these hierarchies have no more than 2 to 3 levels, as the effect becomes less important (Yang, 2006). An example of a hierarchical Bayesian prior will be presented in Subsection 3.2.1 when discussing the hierarchical Bayesian phylogenetic model proposed by Suchard *et al.* (2003).

Markov Chain Monte Carlo (MCMC)

Bayesian phylogenetic models can involve hundreds or thousands of parameters and high dimensional integrals. The calculation of the posterior probability of a phylogenetic tree involves evaluating the marginal probability of the data $p(D)$, which is the sum over all possible tree topologies and integration over all branch lengths in those trees and over all parameters in the substitution model which is hard to calculate analytically (Yang, 2006). Thus, except for trivial problems, a numerical method is needed and the development of MCMC algorithms has resulted in a powerful solution.

The basic idea of MCMC is to construct a Markov chain that has the parameters of the statistical model as its state space and a stationary distribution that is the posterior probability distribution of the parameters. MCMC methods sample successively from a target distribution, which is the posterior. Each sample depends on the previous one, hence the notion of the Markov chain.

Formally, a first-order Markov chain is a sequence of random variables, $\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}, \ldots$ for which the distribution of random variable $\boldsymbol{\Psi}^{(t)}$ depends only on $\boldsymbol{\Psi}^{(t-1)}$. Markov chains have the property that they often converge towards the stationary distribution regardless of the starting point. Different methods can be used to build a Markov chain with a specific stationary distribution and for parameter-rich models usually a mixture of different samplers is typically used, with each sampler targeting one parameter or a set of related parameters. The algorithm can either cycle through the samplers systematically or choose among them randomly according to some proposal probabilities. Together these samplers determine the transition kernel $p(\boldsymbol{\Psi}^{(t+1)}|\boldsymbol{\Psi}^{(t)})$. The most common is known as the *Metropolis algorithm*, originally described by Metropolis *et al.* (1953). Hastings (1970) later introduced an important extension (to be defined on page 33), and the sampler is often referred to as the Metropolis-Hastings algorithm.

In Bayesian inference, instead of searching for the optimal tree, one samples trees according to their posterior probabilities. Once such a posterior sample is available, features that are common among the trees can be discerned (Huelsenbeck *et al.*, 2001).

Computing posterior expectations

MCMC algorithms are closely related with Monte Carlo integration, a simulation method for calculating multidimensional integrals. Let $\boldsymbol{\vartheta}$ denote all the model parameters. The expectation of $h(\boldsymbol{\vartheta})$ over the density $p(\boldsymbol{\vartheta})$,

$$I = E_p[h(\boldsymbol{\vartheta})] = \int h(\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}, \tag{2.18}$$

can be estimated using independent samples $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \ldots, \boldsymbol{\vartheta}_{n_\vartheta}$ drawn from $p(\boldsymbol{\vartheta})$, with I estimated by

$$\hat{I} = \frac{1}{n_\vartheta}\sum_{i=1}^{n_\vartheta} h(\boldsymbol{\vartheta}_i). \tag{2.19}$$

Using the Central Limit Theorem, $\hat{I}$ has an asymptotic normal distribution with mean I and variance

$$\mathrm{var}(\hat{I}) = \frac{1}{n_\vartheta^2}\sum_{i=1}^{n_\vartheta}(h(\boldsymbol{\vartheta}_i) - \hat{I})^2. \tag{2.20}$$

Unfortunately when sampling $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \ldots, \boldsymbol{\vartheta}_{n_\vartheta}$ from an MCMC algorithm, the $\boldsymbol{\vartheta}_i$ form a dependent sample from the target distribution $p(\boldsymbol{\vartheta})$. This dependence affects the calculation of the variance of the estimator $\hat{I}$. Suppose δ_k is the autocorrelation of $h(\boldsymbol{\vartheta}_i)$ over the Markov chain at lag k. Then the variance can be written as

$$\mathrm{var}(\hat{I}) = \frac{1}{n_\vartheta^2}\sum_{i=1}^{n_\vartheta}(h(\boldsymbol{\vartheta}_i) - \hat{I})^2 \times [1 + 2(\delta_1 + \delta_2 + \delta_3 + \ldots)]. \tag{2.21}$$

In effect, a dependent sample of size n_ϑ contains as much information as an independent sample of size $n_\vartheta/[1 + 2(\delta_1 + \delta_2 + \delta_3 + \ldots)]$, known as the *effective sample size* (Yang, 2006).

Metropolis-Hastings

In Bayesian inference the target distribution is the posterior, $p(\boldsymbol{\vartheta}) = p(\boldsymbol{\Psi}|D)$, so that MCMC algorithms generally generate dependent samples from the posterior. The idea of the Metropolis-Hastings (M-H) algorithm is to start the chain at an arbitrary point and in each iteration random changes are performed to current parameter values, and then accepted or rejected according to appropriate probabilities. Formally, the algorithm can be described as follows.

1. Set initial state $\boldsymbol{\Psi}^0$.
2. At t^{th} step (for $t = 0, 1, ...$):
   (a) Propose a new state $\boldsymbol{\Psi}^*$ from *proposal density* $q(\boldsymbol{\Psi}^*|\boldsymbol{\Psi}^t)$.
   (b) Accept $\boldsymbol{\Psi}^*$ with probability $min(1, A)$ where

$$\begin{aligned} A &= \frac{p(\boldsymbol{\Psi}^*)}{p(\boldsymbol{\Psi}^t)} \times \frac{p(D|\boldsymbol{\Psi}^*)}{p(D|\boldsymbol{\Psi}^t)} \times \frac{q(\boldsymbol{\Psi}^t|\boldsymbol{\Psi}^*)}{q(\boldsymbol{\Psi}^*|\boldsymbol{\Psi}^t)} \\ &= \text{prior ratio} \times \text{likelihood ratio} \times \text{proposal ratio}. \end{aligned} \tag{2.22}$$

   (c) If the proposal is accepted, set $\boldsymbol{\Psi}^{(t+1)} = \boldsymbol{\Psi}^*$. Otherwise set $\boldsymbol{\Psi}^{(t+1)} = \boldsymbol{\Psi}^t$.
3. Repeat step 2.

This algorithm generates a random sequence of visited states which constitutes a Markov chain - given the current state, the next state to be sampled does not depend on past states. Also, the proposal density can be symmetrical ($q(\boldsymbol{\Psi}^*|\boldsymbol{\Psi}) = q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^*)$) or asymmetrical ($q(\boldsymbol{\Psi}^*|\boldsymbol{\Psi}) \neq q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^*)$) for any $\boldsymbol{\Psi} \neq \boldsymbol{\Psi}^*$. The algorithm proposed by Metropolis *et al.* (1953) assumes symmetrical proposals and it was extended by Hastings (1970) to allow asymmetrical proposal densities.

Often, when sampling from the posterior, samples are retained only every certain number of iterations, a process that is called *thinning* the chain, as this contributes to reduce the autocorrelation across iterations, disk usage and the output size, easing up further processing. The efficiency of the algorithm can be immensely affected by the nature of the proposal: hence the need of developing efficient proposal algorithms.

Although the prior and likelihood are directly related with the problem in hand, the proposal ratio is solely dependent on the proposal algorithm. The proposal density $q(\cdot|\cdot)$ needs to specify an aperiodic and irreducible chain, i.e the state has period $d = 1$ and the chain is allowed to reach any state from another state, to guarantee that the MCMC algorithm will converge. Each of the proposal mechanisms used in the Bayesian inference framework applied to our proposed model will be discussed in Chapter 3.

Gibbs Sampler

Special cases of the M-H algorithm have been developed such as the Gibbs sampler (Geman & Geman, 1984). The Geman brothers named the algorithm after the physicist J. W. Gibbs, some eight decades after his death, in reference to an analogy between the sampling algorithm and statistical physics. They introduced Gibbs sampling in the context of image restoration. It became popular for Bayesian inference, though it requires conditional sampling of conjugate distributions (see Subsection 2.2.2).

A Gibbs sampler generates a draw from the distribution of each parameter or variable in turn, conditional on the current values of the other parameters or variables. In this sampler the proposal distribution for updating a specific parameter is the conditional distribution of that parameter given all other parameters which leads to accepting all proposals with probability one.

Metropolis-coupled MCMC (MC^3)

The Metropolis-Coupled Markov Chain Monte Carlo (MCMCMC/MC^3) algorithm was introduced by Geyer (1991) for multimodal distributions. The MC^3

method is usually helpful when the target distribution has multiple peaks, separated by deep valleys, which makes it harder for the chain to jump from one peak to another. The strategy is to run m chains in parallel, with different stationary distributions $p_j(\cdot)$, for $j = 1, 2, ..., m$, where $p_1(\cdot)$ is the target distribution. The first chain is the only one that converges to the correct posterior distribution and is known as the *cold chain*. For the other chains (*hot chains*), the stationary distributions will result from an incremental 'heating' such that

$$p_j(\boldsymbol{\Psi}) \propto p_1(\boldsymbol{\Psi})^{1/[1+\lambda(j-1)]}, \text{ for } \lambda > 0. \tag{2.23}$$

By raising the density $p(\cdot)$ to the power $1/T$, with $T = 1 + \lambda(j-1) > 1$, the distribution will be flattened allowing the algorithm to access other peaks. This is accomplished by proposing a swap of states between two randomly chosen chains through a M-H step. Considering $\boldsymbol{\Psi}_j$ the current state in chain j, for $j = 1, 2, ..., m$, the swap between chains j and i will be accepted with probability

$$\min\left\{1, \frac{p_i(\boldsymbol{\Psi}_j)p_j(\boldsymbol{\Psi}_i)}{p_i(\boldsymbol{\Psi}_i)p_j(\boldsymbol{\Psi}_j)}\right\}, \tag{2.24}$$

leading to better mixing. Only the output from the cold chain will be used at the end. Despite having several chains running, only one leads to samples from the posterior, which turns out to be computationally expensive. It is advantageous if the algorithm is implemented to take advantage of parallel processes (Yang, 2006).

Convergence

When the initial state of a chain is far from the posterior mode, the initial likelihood will probably be low. After a long number of iterations, the chain should start moving towards the posterior regions with high probability mass and consequently the posterior density will increase. The samples obtained in this early phase of the run are known as the *burn in*, and are often discarded due to being heavily influenced by the starting point. This is the period in which the chain converges to its stationary distribution.

When the posterior values mix around a ‘plateau’, it might indicate that the chain converged onto the target distribution. Therefore, the plot of the MCMC sampled values or of the overall log-likelihood against the generations of the chain, known as the *trace plot* (Figure 2.5), is often used in monitoring the performance of an MCMC run.

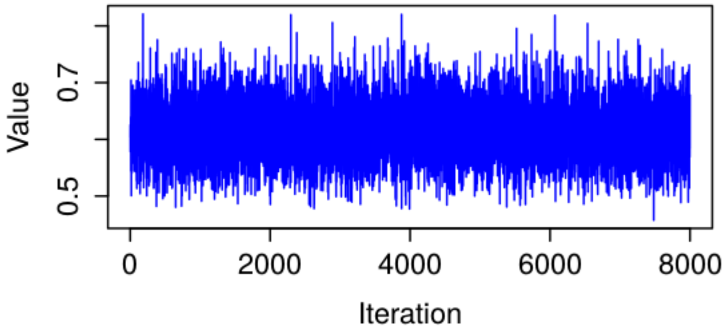


Figure 2.5: Example of a traceplot of the MCMC generated values for a specific parameter.

The decision whether to use a single or multiple runs, using the same model for the same dataset, is related to the ability to diagnose convergence. We have seen that the algorithms used in stochastic tree search might spend long periods in a relatively small region or local peak, which can mislead one into believing that the chain has converged. With many parallel chains it is unlikely that all runs will be showing this behaviour together. However, it is very important to confirm convergence using other diagnostic tools since it is not sufficient for the chain to reach the region of high probability in the posterior, it must also cover this region adequately. The speed with which the chain covers the interesting regions of the posterior is known as its *mixing behaviour*. The better the mixing, the faster the chain will generate an adequate sample of the posterior (Ronquist *et al.*, 2009).

However, the development of correct and efficient MCMC algorithm proposals is a challenging task, especially when dealing with sophisticated parameter-rich models on real data, which usually cause problems for both inference and computation. There might be a lack of information for estimating the multiple parameters, which leads to a nearly flat or ridged likelihood surface or strong correlation between parameters. Also, it is often impossible to calculate an independent posterior distribution to validate the computational implementation. Despite being correctly implemented, an MCMC algorithm can suffer from slow convergence and poor mixing, which causes not only a long time for achieving stationarity but also the sample states might be highly correlated over the iterations resulting in an inefficient parameter space exploration (Yang, 2006).

The proposal mechanism also affects convergence since it needs to facilitate the chain to go from any point in parameter space to any other point in a finite number of steps (with positive probability). But in practice, some proposal mechanisms mix and/or converge much faster than others. This might be due to how the proposal density is specified. It is possible to determine how bold the proposals are by changing the proposal mechanism's tuning parameters. Adjusting the tuning parameter values to reach a target acceptance rate can be done manually or automatically using adaptive tuning methods. Modest proposals will be accepted most of the time taking a longer time to cover the region of high probability mass in the posterior distribution. When too bold, most proposals will be rejected resulting again in a long time to cover the region of interest. Extreme acceptance rates thus indicate that sampling efficiency might be improved by adjusting proposal tuning parameters.

Studies by Roberts *et al.* (1997) and Roberts & Rosenthal (1998, 2001) on several types of complex unimodal posterior distributions suggest that the optimal acceptance rate is 0.44 for one-dimensional and 0.23 for multi-dimensional proposals. Multimodal posteriors are expected to have even lower optimal acceptance rates. Some samplers used in Bayesian MCMC phylogenetics, e.g. some tree topology update mechanisms, have acceptance rates that will remain low, independent of the tuning parameters' values (Ronquist *et al.*, 2009).

Fortunately, several convergence diagnostics have been developed and can help determine the quality of a sample from the posterior distribution. Several heuristic methods have been suggested as diagnostic tools for an MCMC run such as:

- checking the trace plots for all the parameters (convergence of the model parameters is only achieved when all parameters did converge);
- multiple chains started at different points in the parameter space should converge to the same stationary distribution (this approach is arguably the most powerful way of detecting convergence problems, its drawback being the wasted computational power by generating several independent runs);
- run the chain without data, which should lead to a posterior distribution which is the same as the prior;
- generate data under the likelihood model using prior-sampled parameter values (e.g. when assuming that, for a continuous parameter, the $(1-\alpha)100\%$ posterior credible interval (CI) contains the true parameter value with probability $(1-\alpha)$, it is possible to check whether the true value is included in the interval) (Yang, 2006).

When applying Bayesian MCMC methods to phylogenetic problems, usually the most difficult parameter to sample from is the tree topology, making it the key parameter to monitor when checking convergence. By using randomly chosen trees as initial states, the several parallel runs will initially sample from very different regions of tree space. If the chains are converging to the posterior distribution, we expect the tree samples to become more and more similar and therefore, comparing the variance among and within tree samples from different runs seems an efficient convergence diagnostic. The most common approach to compare samples of tree topologies is to focus on split frequencies. Since each branch in a tree corresponds to exactly one split, if two tree samples are similar, the split frequencies should be similar as well. An overall measure of the similarity of two or more tree samples can be the average standard deviation of the split frequencies (used in MrBayes (Huelsenbeck *et al.*, 2012)) or the maximal frequency difference among all observed splits (used in PhyloBayes (Lartillot & Philippe, 2004)). While the former assumes convergence if the average standard deviation is ≤ 0.01, the latter states that a chain has converged once the maximal frequency difference among all observed splits is ≤ 0.1, although it accepts that $0.1 <$ maximal difference among all observed splits < 0.3 results in acceptable convergence. As the tree samples become more similar, both values should tend to zero.

3

Phylogenetic Incongruence

A central premise of phylogenetic analysis is that relationships among organisms follow a hierarchical pattern which can be inferred by observing and analysing homologous traits, or genetically determined characteristics, shaped by evolutionary history. An homologous trait in two species is a trait inherited from their common ancestor. This definition requires, however, an underlying species phylogeny, which itself is a hypothesis and is usually unknown (Dávalos *et al.*, 2012).

Most traits in an organism are expected to have a common evolutionary history but incongruent gene trees might result as a consequence of different rates of change and evolutionary mechanisms (Bull *et al.*, 1993). Incongruence has also become a more detectable problem with the advent of genome-scale data sets. Although highly advantageous in phylogenetic reconstruction, genetic sequences are not without their problems. Not only paralogous sequences (i.e. sequences that diverged after a duplication event within a genome) and sequence alignment are potentially problematic, phylogeneticists are confronted with the dilemma of how to incorporate information about the available multiple genes in their analyses (Cranston *et al.*, 2009; Galtier & Daubin, 2008). Incongruence not only undermines the reconstruction of the underlying species tree from a set of gene trees but also raises the question of the extent to which the history of organisms' evolution can be represented by a single phylogenetic tree (Beiko *et al.*, 2008). Several studies on large data sets have confirmed that phylogenetic conflict is common, and frequently the norm rather than the exception (Dávalos *et al.*, 2012).

Taxonomic sampling (Graybeal, 1998), the number of characters sampled (Rosenberg *et al.*, 2002), and method of analysis (Felsenstein, 1978) can all affect estimates of phylogeny. Adaptive evolution leading to convergence, once thought to be extremely rare (Patterson, 1988), is also a relatively common source of conflict among gene trees, as it is between morphological and molecular phylogenies. Large data sets have also helped establish that high rates of change leading to saturation are common (Dávalos *et al.*, 2012), as well as biological processes leading to different gene trees such as the mechanisms behind *reticulate evolution* (Linder *et al.*, 2004; Bapteste *et al.*, 2005; Degnan & Rosenberg, 2006).

This chapter will explore the phylogenetic incongruence resultant from reticulate evolution, describing in more detail the biological background of lateral gene events and how it affects species tree inference. The hierarchical phylogenetic model for multi gene data of Suchard *et al.* (2003) and its extension to account for LGT events (Suchard, 2005), are also described as they are the basis for the model we are introducing in this thesis.

3.1 Background

When two or more independent evolutionary lineages undergo some type of genetic combination or exchange, we are in the presence of reticulation. It can occur at the chromosomal, population or species levels (Linder *et al.*, 2004). *Hybridisation* (two species recombine originating another species) and *lateral gene transfer* (LGT), when genes are transferred across species boundaries, are the main sources of reticulate evolution at the species level. In the presence of hybridization, no single tree will represent adequately the evolution of the taxa under study, and a network or a set of gene trees will usually be a more appropriate representation. Whether a species tree can be reconstructed from gene trees, if genes are randomly transferred between the trees' lineages, is still controversial and we will return to this question later in this thesis. At the population level (within each lineage) sexual recombination will cause reticulate evolution and shuffling of genes will result from meiotic recombination, which happens at the chromosome level. The diagram in Figure 3.1 published in Linder *et al.* (2004) gives an interesting illustration of these possible reticulation scenarios.

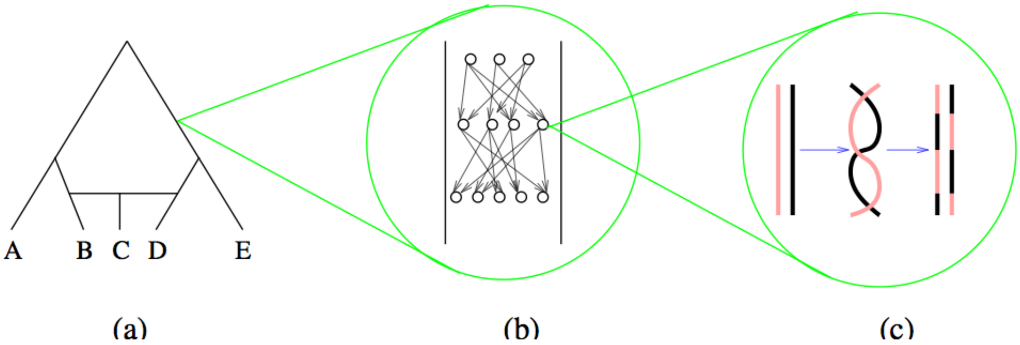


Figure 3.1: **Reticulation events:** a) at the species level representing a hybridisation event; b) at the population level representing parental recombination of genes and c) at the chromosomal level representing meiotic recombination (Linder *et al.*, 2004).

Note that when the aim of the study is species-level inference we might feel inclined to assume that only reticulate events at the species level will make the graphical representation of evolution as a tree-like graph challenging. But although meiotic recombination does not cause a species-level reticulate evolutionary history it may produce patterns that confound species-level inference of reticulation (Linder *et al.*, 2004). In what follows, we will assume that the individual gene datasets are recombination-free (so that meiotic recombination, or exchanges between sister chromosomes, does not take place), thus simplifying our analysis and allowing us to assume that all gene evolution is tree-like (Posada & Wiuf, 2003; Zhang & Jin, 2003). As meiotic and sexual recombination are not in the scope of our work we are redirecting our attention to reticulation events among species.

In hybrid speciation, two lineages recombine to create a new species, as symbolized in Figure 3.2(a), but the evolutionary history of the genes inherited from species X and Y can still be individually represented by a gene tree (see Figures 3.2(b) and 3.2(c)). In the presence of an LGT, although genetic material is transferred from one lineage to another it does not necessarily result in a new lineage (Figure 3.2(d)), so that the evolution of a gene can still be represented as a tree. For this specific case, the tree in Figure 3.2(e) represents the evolutionary history of the genes laterally transferred from another species, and in Figure 3.2(f), the genes inherited from the parent. Therefore, although trees might not be the most appropriate graphical models of species evolution when reticulation occurs, they are still appropriate for gene evolution.

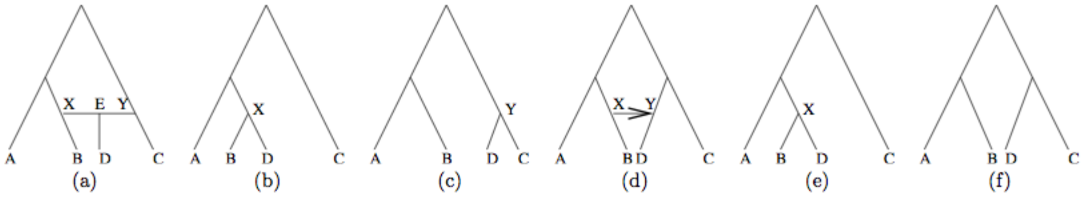


Figure 3.2: **Hybridisation and LGT:** a) Species D is a hybrid of species X and Y. Evolutionary history of genes inherited from species X can be represented by gene tree (b) whereas tree (c) represents genes inherited from species Y. A similar scenario is produced in the presence of an LGT represented in subfigures (d), (e) and (f) (Linder *et al.*, 2004).

Both types of reticulation event are sufficiently common to be of serious concern to systematists. Hybrid speciation is common in some very large groups of organisms: plants, fish, amphibians, and many lineages of invertebrates, and horizontal gene transfer appears to be very common in bacteria with lower levels being evident in many multicellular groups.

This thesis focuses on LGT driven reticulate evolution. Thus, we will now explore, in more detail, the most important biological aspects of gene transfers between organisms and species tree reconstruction in their presence.

3.1.1 Lateral Gene Transfer

Lateral gene transfer, 'the non-genealogical transmission of genetic material from one organism to another' (Goldenfeld and Woese 2007) is widely accepted as a source of new genes and functions to the organisms that received the genetic material, assuming the survival of that material throughout the subsequent generations.

LGT is most likely to occur between closely related species, but can also occur between distantly related organisms. A considerable number of published studies about genes that have probably been acquired by LGT show that the transfer can occur not only within domains such as from Bacteria to Bacteria (Ochman *et al.*, 2000; Ku *et al.*, 2013), Archaea to Archaea (Doolittle & Logsdon, 1998; Kaminski *et al.*, 2013) and Eukaryotes to Eukaryotes (Andersson, 2005; Wisecaver *et al.*, 2013) but also between domains in all possible directions (Boto, 2010; Nikolaidis *et al.*, 2013; Yue *et al.*, 2013; Robinson *et al.*, 2013). The length of DNA segments believed to have been laterally transferred seem to range from 7 nucleotides (Denamur *et al.*, 2000) to an entire chromosome greater than 3 Mb (Lin *et al.*, 2008). These segments include non-coding DNA, portions of genes, intact genes, multi-gene clusters, operons, plasmids, transposable elements and pathogenicity islands, hence the proposal to use the expression 'lateral *genetic* transfer' rather than *gene* (Beiko *et al.*, 2005).

Mechanisms behind the transfer of genetic material between micro-organisms became well known in the early research stages in molecular biology and molecular genetics (Lederberg & Tatum, 1946; Zinder & Lederberg, 1952; Stocker *et al.*, 1953). In a nutshell, the foreign genetic material enters the cell, either as a naked sequence or in a vector (e.g. plasmids, integrons, transposons) and once inside, the gene must escape the host defenses, be incorporated into the host genome and become expressed as a functional protein. This new gene will only be maintained if it provides a function which is selected for in the population and it may or may not replace homologous genetic material. Basic lateral transfer mechanisms include transformation, conjugation-mediated plasmid exchange, phage-mediated transduction and variations of these processes (Figure 3.3). In contrast to transformation and phage-mediated LGT, conjugation requires physical contact of the donor and recipient bacteria. Genetic integration into a host genome might result from homologous recombination, illegitimate recombination, combinations of both these mechanisms or site-specific recombination (Brigulla & Wackernagel, 2010).

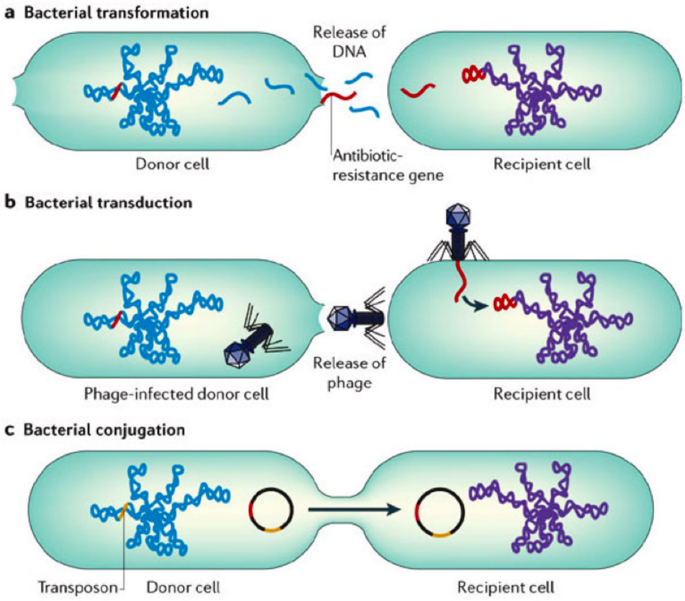


Figure 3.3: **Mechanisms of lateral gene transfer (LGT) in bacteria:** a) Transformation occurs when naked genetic material is released on lysis of an organism and is taken up by another organism; b) In transduction, genes are transferred from one bacterium to another by means of phages and can be integrated into the chromosome of the recipient cell (lysogeny); c) Conjugation occurs by direct contact between two bacteria: plasmids form a mating bridge across the bacteria and genetic material is exchanged (Furuya & Lowy, 2006).

The success of any LGT attempt depends on the vector compatibility between donor and recipient, which is often related with recognition of, and interaction with, recipient surface proteins. Despite the limitations, conjugation between distantly related organisms such as bacteria and eukaryotes has been demonstrated experimentally (Beiko *et al.*, 2005).

As a result of these findings, Syvanen proposed in 1985 the theoretical effect upon evolution of gene transfer across species. The theory suggests that genes can be transferred and expressed among all species and that the uniformity of the genetic code would allow organisms to decipher and use genes transposed from chromosomes of foreign species. But it was only when phylogeneticists used the sequence of 16S RNA genes for reconstructing old phylogenetic relationships that they realized that these genes were grouping together micro-organism species that were split by other morphological, physiological or molecular markers (Boto, 2010). In 1993, Hilario and Gogarten proposed the concept of LGT between organisms as an alternative explanation for the observed phylogenetic conflicts. Since then, especially with the rise of the genomic era which allowed the comparison of complete sets of genes between organisms, new and abundant data have reinforced this idea.

A great effort has also been carried out in the past years to gain an insight on the importance of LGT events in Bacterial and Archaeal evolution (Boto, 2010). Nonetheless the results are controversial since any topology obtained in phylogenetic studies that supports an LGT event may also be explained by gene duplication and gene loss events (Kurland *et al.*, 2003). In order to distinguish between them, the likelihoods of these evolutionary events have to be considered, which is problematic given that the frequencies of such events are expected to vary between lineages as well as throughout evolution (Andersson, 2005). Nevertheless, it seems that lateral genetic transfer does play a larger role in microbial evolution than initially thought (Dagan *et al.*, 2008), and despite the fact that the number of fully sequenced genomes available has risen dramatically in past years, the bacterial phylogeny seems to be still largely unresolved. These facts led to the proposal that, in the presence of such a high rate of LGT events in bacterial evolution, the species tree can no longer be recovered (Doolitle & Bapteste, 2007), although more conservative opinions exist. Galtier & Daubin (2008) showed that, although LGT significantly influences the bacterial phylogenomic pattern, it probably still allows the reconstruction of the species tree in this group. They argue that 'phylogenetic agreement is much more common than disagreement, indicating that LGT is not prevalent enough to erase the vertical signal'. More recently, Roch & Snir (2013) claimed that, under a model of randomly distributed LGT, the species phylogeny can be reconstructed even in the presence of many LGT events per gene tree.

With regard to eukaryotic evolution, although it has been assumed that the lateral gene transfer effect is less relevant, it does not seem to be a negligible evolutionary force both in unicellular (Andersson, 2005; Keeling & Palmer, 2008) and multicellular (Boto, 2010) eukaryotes. Independent of the organism, if genetic exchanges occur between species, then the phylogeny of individual genes will be influenced by the number and nature of transfers they have undergone (Galtier & Daubin, 2008). In general, only a small number of genes are expected to have been horizontally transferred between any given pair of species, although the concept of a *highway of gene sharing* has been proposed by Beiko *et al.* (2005) to represent the multitude of LGT events that happen between some pairs of species.

Two general methods have been proposed to examine LGT. First, by examining variation in nucleotide base composition and bias in codon usage in single genomes, genes suspected to have been imported through LGT are potentially identifiable. The other method uses comparative studies across species and is based on using phylogenetic incongruence to reconstruct species-level networks or phylogenetic trees. This last method offers the advantage of having a direct biological interpretability as it describes the underlying evolutionary histories of the different genes (Suchard, 2005).

Networks

Recently, there has been some interest in using networks rather than trees to represent evolutionary relationships between species that have undergone reticulate evolution or to represent conflicts with a treelike evolutionary framework (Huson *et al.*, 2011). Phylogenetic networks generalize evolutionary trees, and can represent evolutionary histories of species that have undergone reticulate evolutionary processes such as hybridization, recombination and lateral gene transfer. Huson & Bryant (2006) define a phylogenetic network as any network in which taxa are represented by nodes and their evolutionary relationships are represented by edges. This definition allows the classification of phylogenetic networks in different types including phylogenetic trees, split networks, reticulate networks and other type of networks representing evolutionary data (see Figure 3.4). In a general sense, most networks allow the graphical representation of evolutionary events not only where species speciate but also combine (Iersel & Moulton, 2014).

A split network, widely used as a visualisation tool of potential phylogenetic conflicts, is obtained as a combinatorial generalisation of phylogenetic trees representing incompatibilities within and between datasets and was proposed by Huson & Bryant (2006) also as a statistical inference tool.

Reticulate networks are used to represent evolutionary histories in the presence

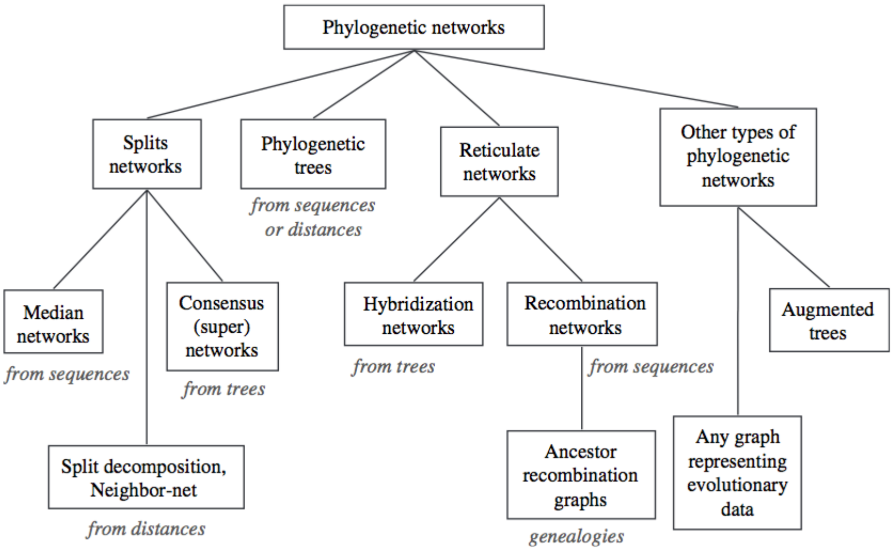


Figure 3.4: A diagram representing the phylogenetic networks classification, according to Huson & Bryant (2006), which includes a number of different concepts such as phylogenetic trees, split networks and reticulate evolution.

of reticulate events. They usually represent reticulate evolution as a phylogenetic tree with additional edges where a node with two or more ancestors corresponds to a reticulate event. Reticulate networks are usually rooted giving a time direction with an evolutionary meaning and can be split into two types: hybridisation networks and recombination networks. The former explains a given set of trees in terms of hybridisation events and the aim is to determine a putative reticulate network $\mathcal{N}$, given from which trees in $\mathcal{T}$ will arise (Huson & Bryant, 2006). As an example, Maddison (1997) reconstructs such a network by first inferring individual gene trees from separate analyses and then reconciling the trees into a network. Maddison observed that when there is one reticulation in the network, there are two trees within the network, and every gene evolves according to one of these two gene trees. More generally, Maddison suggested that a network that contains multiple reticulations can be reconstructed from its constituent gene trees. Given two gene trees, one can reconstruct a network with the smallest number of reticulations which induces both trees.

Linder *et al.* (2004) suggested a classification of approaches for phylogenetic network reconstruction into several categories. The first one will originate a phylogenetic tree and suggests that, when the reticulate event originating the network is an LGT, a possible approach is to identify the genes involved in it and remove them from the analysis, reconstructing a tree based on the remaining genes. The second approach starts by building a tree and adding non-tree edges using a greedy approach to optimise some cost criterion (Hallett & Lagergren, 2001; Clement *et al.*, 2000; Makarenkov, 2001). A third approach attempts to reconcile several trees built with different subsets of the data and underpins the median networks (Bandelt *et al.*, 2000) and the molecular-variance parsimony approach (Excoffier *et al.*, 1992). The parts of the tree where the reconciliation fails might be explained by a reticulation event.

This thesis focuses specifically on the reconstruction of gene trees given a phylogenetic tree that represents the evolution of a number of species in the presence of gene transfer events, and therefore we will narrow our attention to phylogenetic tree reconstruction.

3.2 Phylogenetic Species Tree Reconstruction

If a reconstructed gene tree differs from the assumed phylogeny of the species being studied, then LGT might be a possible explanation. Nevertheless, the true species tree is often unknown making it necessary to either fix the species tree to equal an inferred tree for a specially chosen gene (usually highly conserved among species) or simultaneously estimate the species tree and gene trees given a biologically plausible model relating them (Suchard, 2005).

To address this issue, two main groups of approaches have been proposed: a strict combined-data approach where multiple genes are combined into a single undifferentiated partition before phylogenetic analysis, and the consensus approach which has a two stage methodology: fit an independent phylogenetic model to each gene to estimate separate evolutionary histories and reach a consensus tree from the resulting gene topologies.

Kluge & Wolf (1993) claim that natural data partitions do not exist and the species tree should be estimated using the whole sequence of the genome. They proposed a combined-data approach in which the sequences from all available genes are concatenated into a single sequence, along with other phylogenetic characters such as morphology or behaviour.

An apparent advantage of this approach is that parameter estimates and inference regarding the single evolutionary pattern are more robust than those from individual genes. Data for individual genes are potentially sparse and will be more subject to the effects of sampling variation (Suchard *et al.*, 2003). On the other hand, this method ignores the existence of the gene as the basic functional unit of the genome, which has drawn criticism (Slowinski & Page, 1999) given it assumes that the simultaneous analysis approach erroneously treats every nucleotide of all available genes as independent estimators of the underlying species phylogeny, which would mean that the longer the sequence the more precise the estimated species tree. The estimate of the species tree is then biased if the gene trees for the long genes happen to have incorrect topologies. Also, it is now generally known that gene trees in principle may not match the species tree irrespective of whether the gene has a long sequence or a short sequence. Indeed, it has been shown that under some combinations of branch lengths in the species tree, incongruent gene trees are more likely to occur than congruent gene trees (Kubatko *et al.*, 2007; Degnan & Rosenberg, 2006). Another disadvantage of this approach lies in the fact that it ignores single gene information implying that, given the evolutionary history for the concatenated data, results from the individual genes are irrelevant. But as we have discussed before, individual genes may evolve at different rates, under different pressures or have been laterally transferred and separate analysis can give different insights into the histories of each gene. This will result in assuming different evolutionary models for one or more genes to avoid inconsistent or biased inference (Yang, 1995; Buckley *et al.*, 2002).

For the consensus tree method, gene trees are inferred separately for each gene, and the consensus tree, i.e., an agreement tree between two or more trees of these gene phylogenies, is used as the estimate of the species tree. It summarizes congruence among individual gene trees and produces high resolution in the branching pattern only when there is at least a majority consensus among the different data sets (Gadagkar *et al.*, 2005). The argument in favour of the consensus approach includes the fact that it accounts for extensive differences in evolutionary rates and substitution patterns among genes in a gene specific manner. Nevertheless, retrieving only one topology means that uncertainty is not accounted for. Also, the process for estimating gene trees and for estimating species trees should not be independent. Gene trees for different genes are dependent since they all depend on the species tree. Consequently, it is more appropriate to assume only conditional independence of the gene trees given a common species tree. According to this assumption, the gene trees should then be estimated jointly across multiple loci. An extension of the first level of a consensus approach into a Bayesian framework was proposed by Buckley *et al.* (2002) with the purpose of analysing the congruence of gene tree topologies. But, as in the consensus analysis, it inferred individual gene parameters independently and then averaged these values or compared them with a fixed point (Suchard *et al.*, 2003). As a result, it fails in accounting for dependency among the gene trees.

To overcome the inherent difficulties of both approaches, a mixture of them is often employed by dividing the parameter space into two sets: one set of parameters is fixed across genes while the other parameters remain conditionally independent between genes (Yang, 1996). This type of analysis is present in popular phylogenetic software such as MrBayes (Huelsenbeck & Ronquist, 2001), PAML (Yang, 1997), PAUP* (Swofford, 2003) and Phylip (Felsenstein, 1989).

3.2.1 Suchard's Bayesian Hierarchical Phylogenetic Model for Multiple Gene Data

Bayesian hierarchical models can be used as an alternative in analysing multiple genes due to them naturally averaging across uncertain discrete quantities such as topologies across genes. In 2003, Suchard *et al.* proposed a Bayesian hierarchical phylogenetic model for analysing multiple gene data. The model was later extended (Suchard, 2005) by incorporating the occurrence of lateral gene events between the species in the model. Although our research is based on this latter article, it is important to introduce this hierarchical structure since a similar framework will also be integral to our model.

In Suchard's model all of the data is used in a single analysis as in the combined-data approach but it allows for individual inference of different gene-level phylogenetic parameters as in the consensus approach. A formal statistical model combines the results from the individual genes parameters, including gene topologies, providing across-gene-level summaries of all parameters. When fitted simultaneously with the individual gene models, it enables the borrowing of strength of information from one gene to another in the form of a prior.

To define a hierarchical structure in the model, we start with the natural division of the sequence data $\boldsymbol{D} = (D_1, \ldots, D_g)$ into g separate genes with g copies of the model parameters, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$. This yields

$$p(\boldsymbol{\theta}|\boldsymbol{D}) \propto p(\boldsymbol{\theta})p(\boldsymbol{D}|\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g) \prod_{i=1}^{g} p(D_i|\boldsymbol{\theta}_i). \tag{3.1}$$

Key to Suchard's hierarchical construction is modelling the prior $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ such that it depends on unknown parameters $\boldsymbol{\phi}$ in which $\boldsymbol{\theta}_g$ are only conditionally independent given $\boldsymbol{\phi}$, that is

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g) = \int_{\boldsymbol{\phi}} \prod_{i=1}^{g} p(\boldsymbol{\theta}_i|\boldsymbol{\phi})p(\boldsymbol{\phi})d\boldsymbol{\phi}. \tag{3.2}$$

Employing unknown parameters $\boldsymbol{\phi}$ that, in turn, have their own prior $p(\boldsymbol{\phi})$ enables the borrowing of strength of information from D_g through $\boldsymbol{\theta}_g$ and $\boldsymbol{\phi}$ to the remaining $g-1$ genes and their respective parameters. The hierarchical model thus becomes

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{D}) \propto p(\boldsymbol{\phi}) \prod_{i=1}^{g} p(D_i|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\phi}). \tag{3.3}$$

When assuming gene independence, $\boldsymbol{\phi}$ is a constant across the independent areas of the parameter space. No information is shared and therefore these models are not hierarchical.

A Bayesian hierarchical framework is used to combine the separate gene models within a single comprehensive model which allows information about the values of the parameters in one gene to help in inferring the parameters in other genes. Not only is it useful when some sections of the data are uninformative, but also, these upper-level parameters might reveal tendencies across genes.

Defining the Hierarchical Priors

This model assumes independent and identically distributed sites within a gene. The likelihood of observing D_{gl} is given by a multinomial distribution over the 4^N possible outcomes. The probabilities are functions of an unknown topology τ_g relating the taxa for gene g, the branch lengths $\boldsymbol{\ell}_g$ and a model that describes nucleotide mutation along the branches. Suchard assumes a reversible model for nucleotide substitution, TN93, which is parametrised by the stationary distribution for the nucleotide frequencies and transition:transversion rate ratios for purines υ_g and pyrimidines γ_g.

Branch lengths might not retain their characteristics between topologies, which in their turn might differ across genes. Therefore, in order to be able to share branch length information across genes, each branch length ℓ_{gs} is assumed to be exponentially distributed

$$\ell_{gs} \overset{indep}{\sim} \text{Exp}(\lambda_g) \tag{3.4}$$

with unknown prior expected divergence λ_g for gene g. Given that the likelihood is a negative exponential function of the branch lengths, the exponential distribution is a common choice for a vague prior on branch lengths (Ronquist *et al.*, 2009). The model is also restricted to unrooted trees, since without the molecular clock assumption it is impossible to identify the root (Suchard, 2005).

Expected divergence λ_g and transition:transversion rate ratios υ_g and γ_g exist on the positive half of the real line. Since ratios are naturally transformed onto the entire real line through a logarithmic transformation, they are naturally modelled by a log-normal prior such that

$$\left.\begin{pmatrix} \log \upsilon_g \\ \log \gamma_g \\ \log \lambda_g \end{pmatrix}\right| \boldsymbol{\Theta}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\Theta}, \boldsymbol{\Sigma}) \tag{3.5}$$

where $\boldsymbol{\Theta} = (U, G, M)^t$ are log-scale unknown means and $\boldsymbol{\Sigma} = \text{diag}(\sigma^2_\upsilon, \sigma^2_\gamma, \sigma^2_\lambda)$ is a unknown diagonal variance-covariance matrix. This framework induces correlation between, for example, $\upsilon_1, \ldots, \upsilon_G$, after marginalising over $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$, enabling a sharing of strength of information from one gene to another. However, conditional on $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$, the υ_g, γ_g and λ_g are independent over g.

Conjugate priors are assigned to each upper-level unknown parameter such that

$$\boldsymbol{\Theta} \sim N(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) \qquad \text{and} \qquad 1/\sigma^2_x \sim Gamma(\psi_1, \psi_2), \tag{3.6}$$

for $x \in (\upsilon, \gamma, \lambda)$. This way the computation is facilitated by using direct Gibbs sampling of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$. Suchard chose relatively uninformative priors by setting $\boldsymbol{\Psi}_1 = 0 \times (1, 1, 1)^t$, $\boldsymbol{\Psi}_2 = 10 \times I$, where I is the identity matrix, $\psi_1 = 2.1$ and $\psi_2 = 1.1$. These specifications give a prior expectation of 1 and a variance of 10 for $\sigma^2_\upsilon, \sigma^2_\gamma$ and σ^2_λ.

The Dirichlet distribution is the natural prior for the stationary distributions π_g since they are defined on the unit simplex in $\mathbb{R}^4$. Therefore,

$$\boldsymbol{\pi}_g | N_\Pi, \boldsymbol{\Pi} \sim Dir(N_\Pi \times \boldsymbol{\Pi}) \tag{3.7}$$

where $\boldsymbol{\Pi} = (\Pi_A, \Pi_G, \Pi_C, \Pi_T)$ are the proportions, across-gene level, for each type of nucleotide and N_{Π} is an unknown across-gene level measure of precision for π_g. The priors on $\boldsymbol{\Pi}$ and N_{Π} are assumed to be

$$\boldsymbol{\Pi} \sim Dir(\eta) \qquad \text{and} \qquad N_{\Pi} \sim Gamma(\nu_1, \nu_2), \tag{3.8}$$

where $\eta = (1, 1, 1, 1)$, $\nu_1 = 0.1$ and $\nu_2 = 0.1$, providing a flat prior on $\boldsymbol{\Pi}$ and a proper yet vague prior on N_{Π}.

3.3 Modelling Lateral Gene Transfers

Several research groups have worked on the problem of reconstructing a species tree given gene trees subject to LGT (e.g. the parsimony-based reconciled tree work on Page (2000) and the algorithm developed by Mirkin *et al.* (2003)). Random models for LGT have been studied in a number of papers (Suchard *et al.*, 2005; Galtier, 2007; Linz *et al.*, 2007; Szöllosi *et al.*, 2012; Roch & Snir, 2013; Steel *et al.*, 2013), all assuming that random LGT events occur according to a Poisson process with the rate of transfers between two points in the tree either being constant or being dependent on the phylogenetic distance between the two points. Roch & Snir (2013) showed how a species tree can be reconstructed from a given number of gene trees, provided that the expected number of LGT events lies below a certain threshold. Above this threshold, it becomes impossible to distinguish the underlying species tree from alternative trees.

A class of stochastic models was proposed by Suchard (2005) for LGT that enables the simultaneous estimation of the underlying species tree relating a group of organisms and the gene trees subject to LGT for a set of gene alignments. These models are defined in a hierarchical manner. The hierarchical structure described in Subsection 3.2.1 is used at the within-gene level and across-gene substitution model level, in order to reconstruct each gene tree from its corresponding gene alignment. Simultaneously the LGT models impose an additional structure over the gene tree topologies. The hierarchical model describes the gene trees given an unknown species tree, an unknown number of LGT events and an unknown set of substitution model parameters leading from that species tree to each gene tree. Inference is performed via a Bayesian framework which naturally handles uncertainty in discrete parameters such as the trees and the number of LGT events.

To build a stochastic model for LGT it is essential first to understand the concept of *tree space*, which we addressed in Chapter 2. A more visual approach is to consider tree space as a mathematical graph whose vertices represent all possible trees and the edges describes a direct connection between two vertices. Two vertices that are joined together by a single edge are called *adjacent*, and the set of all adjacent vertices is its *neighbourhood*. Usually a set of tree rearrangement operations is used to define the notion of adjacency, and hence determine the graph. Tree rearrangement operations, which modify the tree topology by applying structural changes, can be used to model the occurrence of an LGT event in a tree topology.

One of the main tools used to understand and model reticulation events is the graph-theoretic operation called *subtree-prune-and-regraft (SPR)*, represented in Figure 2.4(b). Formally, an SPR operation on a phylogenetic tree T is defined as cutting any edge and thereby pruning a subtree t, and then regrafting t by the same cut edge to a new vertex obtained by subdividing a pre-existing edge in T. In this thesis, and in common with Suchard, we assume that any gene analysed has one copy in all the species under analysis and at all speciation times until the most recent common ancestor. Under this assumption, the effect of an SPR on a tree topology is equivalent to a "copy and overwrite" lateral transfer event. Given any two incongruent phylogenetic trees where the incongruence can be explained by a single reticulation event, one tree can be constructed from the other by a single SPR. If more than one reticulation event is needed to explain the incongruence, the events can be modelled by a series of SPRs. Applying an SPR to one topology τ results in the creation of one of several new possible topologies that differ from τ by an extent dependent on the operator. The collection of all trees one operation away from τ becomes its neighbourhood under that operator. Several important properties about the graph induced by the SPR operator in unrooted and rooted trees have been previously studied (Allen & Steel, 2001).

In Suchard's model, LGT is modelled via an unweighted random walk process in the tree space graph, derived from the SPR operator on unrooted phylogenetic trees. One straightforward stochastic process on a graph is an unweighted random walk which proceeds from vertex to vertex along the edges of the graph. This process will generate a discrete-time Markov chain with the visited vertices as the states of the chain. As it is unweighted, the chain randomly chooses the next vertex to visit from all neighbours of its current vertex. For this Markov-chain, the one-event transition probability matrix X is

$$(X)_{uv} = \begin{cases} \dfrac{1}{deg(u)} & \text{if vertices } u \text{ and } v \text{ are adjacent,} \\ 0 & \text{otherwise,} \end{cases}$$

where $deg(u)$ is the degree of u. On the basis of G random walks on the graph induced by the SPR operator, one for each gene, a hierarchical prior is constructed over the joint distribution of all gene trees $\tau_1, \ldots, \tau_G$, assuming that

- the vertex representing the species tree S is the initial state of G Markov chains;

- the Markov chains are conditionally independent given S and X;
- the vertex representing τ_g is the final state of the gth chain; each chain is of unknown length $0 \leq \kappa_g < \infty$.

An example of a set of possible paths of $G = 4$ Markov chains is represented in Figure 3.5.

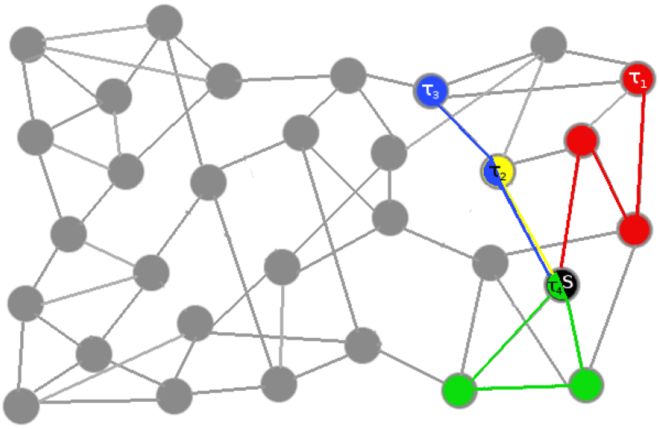


Figure 3.5: **Mathematical graph representing an area of the tree space.** Each vertex corresponds to a possible topology while edges describe a direct connection, related with a specific tree rearrangement operation, between adjacent trees. One possible Markov chain realisation for species tree S and 4 gene topologies $\tau_1, \ldots, \tau_4$. All chains share the starting point S. Chains 1 and 4 have $\kappa_1 = \kappa_4 = 3$, chain 2 has $\kappa_2 = 1$ and chain 3 has $\kappa_3 = 2$.

Given these assumptions, the probability of species tree S giving rise to gene tree τ_g after κ_g LGT events is

$$Pr(\tau_g = v | S = u, \kappa_g) = (X^{\kappa_g})_{uv}. \tag{3.9}$$

Note that this is typically impossible to compute explicitly given the size of the graph. The hierarchical specification is completed by assigning a prior distribution over S by letting

$$S|\mathbf{z}, M \sim Multinomial(1, \mathbf{z})$$

where $\mathbf{z} = (z_1, \ldots, z_M)$ are constants, the prior probabilities of the M possible N-taxon tree topologies. It assumes also a conditionally independent prior on all κ_g such that

$$\kappa_g \overset{indep}{\sim} Poi(\Lambda_g)$$

where Λ_g is the expected number of LGT events for gene g and is a deterministic function of across-gene level parameters. This prior is conjugate to Equation 3.9 allowing all κ_g to be integrated out of the model (see Subsection 2.2.2 for more details on integration). A graphical depiction of Suchard's complete model is shown in Figure 3.6.

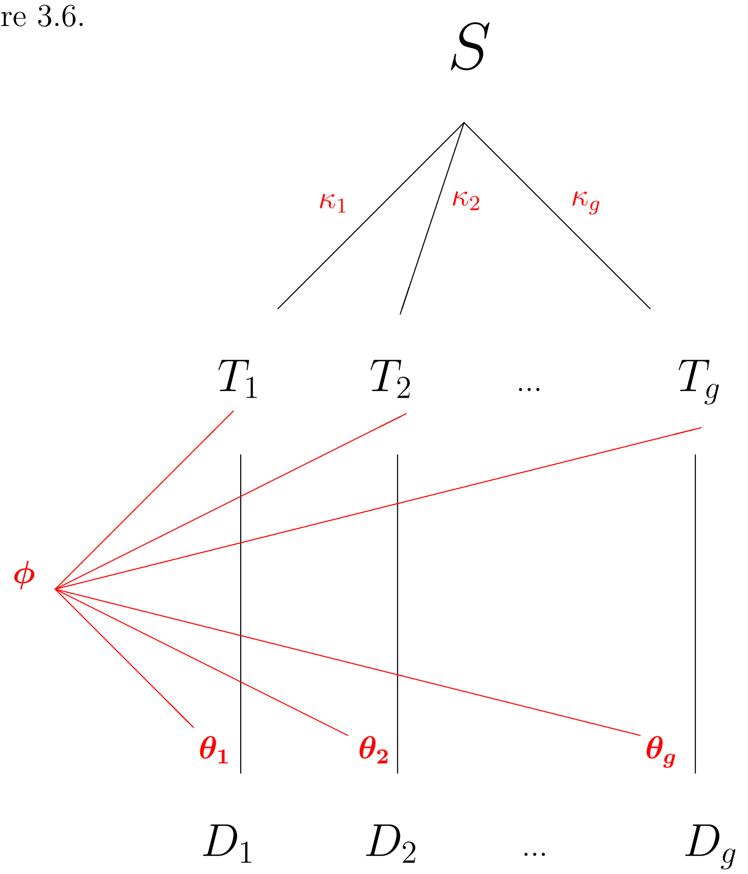


Figure 3.6: **Graphical representation of Suchard *et al.* (2005) hierarchical phylogenetic model.** S represents the species tree, $T_i = (\tau_i, \boldsymbol{\ell}_i)$, for $i = 1, \ldots, g$, is the gene tree for gene i, D_i is the multiple sequence alignment for gene i, θ_i and κ_i are the substitution model parameters and number of LGTs for gene i and ϕ is the hierarchical prior parameter. The edges represent the conditional dependence between parameters.

4

Novel Bayesian Hierarchical Phylogenetic Model for LGT

4.1 Novel Aspects of Our Approach

Our starting point is the Bayesian hierarchical model proposed in Suchard *et al.* (2005) and described in Chapter 3 where a class of stochastic models are used for LGTs that enable the simultaneous estimation of the underlying species tree relating a group of organisms and the gene trees subject to LGTs for a set of gene alignments. This model has several drawbacks related with certain unrealistic biological assumptions. Therefore we propose a more biologically realistic approach to it by introducing the following aspects.

- **An Ordered Rooted Species Tree.** We assume the species tree is a rooted tree (instead of unrooted as previously) and take into account possible time orderings of divergence events in trees, without explicitly modelling divergence times. This corresponds to assigning an order to the internal vertices, in such a way that all speciation events occur at different moments in time.

- **Species Contemporaneity on LGT events.** The time ordering in the previous assumption imposes a natural constraint to LGT events by allowing them to happen only between species that are contemporary.
- **Extended SPR (xSPR).** We use an extended version of the SPR operator (xSPR), which respects the time ordering and describes the effect of an LGT between contemporary species in an ordered rooted tree.
- **Site Evolution Rate Heterogeneity.** We assume that variation in the rate of evolution in different sites is not constant but Gamma distributed.

As we are working under a Bayesian framework, Markov Chain Monte Carlo (MCMC) methodologies were used for parameter inference and two new Bayesian MCMC proposals were developed:

1. **Proposal for ordering a phylogenetic tree.**
2. **Joint proposal for LGT history, LGT distance and gene trees.**

Each of these aspects are described in more detail in the next sections.

In addition we have developed the **LGT Biplot**, a novel visualisation tool for displaying inferred gene transfer history adequately and in an easily understandable way, providing accessible and intuitive means for biologists to explore the results. This tool will be described in more detail in Chapter 5.

4.2 Ordered Rooted Species Trees

A central aspect of organisms' evolutionary histories is the timing of species diversification. Although several widely used models have been proposed to infer speciation times (Drummond & Rambaut, 2007; Liu & Pearl, 2007), in most circumstances any inference must rely almost exclusively on molecular data constrained only by scarce time information (Szöllosi *et al.*, 2012). For this reason, when the study aim is the pattern of evolution rather than timing of events, non-clock-like trees are often used to represent evolutionary paths in which branch lengths represent the expected number of nucleotide substitutions between species. However, the transference of DNA between two organisms implies the contemporaneity of both organisms and in result, modelling LGT events in a non-clock-like tree is problematic.

The effect of an LGT on a rooted species tree topology is represented in Figure 4.1. As we already mentioned in Section 3.3, in this thesis we assume that LGTs are "copy and overwrite" transfer events. At a specific point in time, a gene was transferred from a parent or ancestor of species E, s_E, and an ancestor of species X, s_X. When an LGT occurs between two species, those species will become genetically similar for that specific gene. As a result, when inferring the gene tree, s_X and its descendants will be localised near s_E, the species that donated the gene, even if they are very distant in the species tree. Species s_E will correspond to the internal vertex representing the ancestor of E and X. If a second LGT occurs for that specific gene, time constraints will then affect the possible donors and receivers of that copy of the gene. No species that existed before the time point where the first LGT occurred (in yellow) should be able to transfer genes to the subtree that is moved by the LGT (in blue) given that they are not contemporary; note that the yellow edges represent ancestors of blue edge species.

In order to model ancestor-descendant relationships on a phylogenetic tree, first of all, a time direction must be associated with its edges which can be obtained when assigning a root. In practice, using rooted or unrooted trees leads to observable differences when trying to model LGT or recombination events in the evolutionary histories of a set of species or individuals. For example, in Hein (1993), when propos-

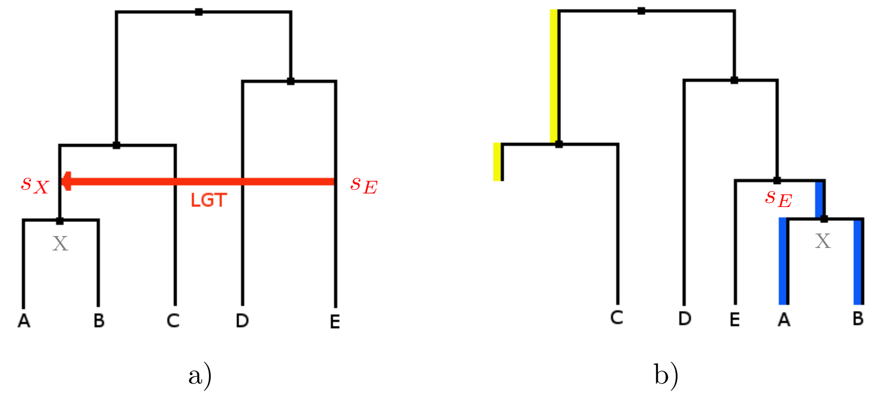


Figure 4.1: **Representation of the effect of LGT on a rooted tree topology.** Assuming that an LGT, represented by the red arrow in a), occurred from species s_E to species s_X, when inferring the corresponding gene tree, the subtree representing s_X and all its descendants is attached below s_E, as represented in b). As a result, time constraints are naturally created for other LGTs. For example, in this case, the species represented in the yellow edges cannot transfer the gene to any species in the blue edges since the yellow edges represent ancestors of blue edges species, which are not contemporary.

ing an algorithm for reconstructing the most parsimonious evolutionary histories of sequences which have undergone recombination (which has the same effect as LGT in phylogenetic trees), if unrooted trees were used in the algorithm, internal contradictions could arise, making it difficult to construct a graphical representation. Although the use of rooted trees would avoid this type of conflict, another problem develops when non-clock-like rooted trees are used instead. Even when two species are not in the same path descending from the root and could be available to transfer genetic material between each other, biological events occur with a certain time ordering. It may happen that actually those two species did not exist at the same time. In this case, by using *ordered rooted trees*, where the information of relative ages of internal vertices is retained, it avoids possible contradictions in representing evolutionary histories of biological sequences (Song, 2003; Szöllosi *et al.*, 2012).

Definition. An ordered rooted tree S_o is a leaf-labelled rooted binary tree whose corresponding set $\{v_1, v_2, \ldots, v_{n-2}\}$ of degree-3 vertices is a totally ordered set defined by age ordering. Computationally the ordering is achieved by assigning a pseudotime $t(\cdot)$, from the interval [0,1], to each vertex such that if u is a descendant of v, then $t(u) < t(v)$. If there exists no ancestral relation between v and u either $t(u) < t(v)$ or $t(u) > t(v)$ is allowed. The pseudotime duration of each edge is determined by the time order of speciation events such that, assuming that vertices are numbered in increasing order backward in time, the pseudotime of the i-th vertex is $t(v_i) = i/W$, where W is the number of internal vertices. Therefore, tree leaves will be assigned pseudotime 0 while $t(v_{root}) = 1$. Any time interval between two successive speciation events is considered a *time epoch* such that $E_i^j = [t(v_i), t(v_j)]$. We further assume equal pseudotime duration of all epochs on S_o. We do not make explicit inferences about the pseudotimes; rather, they are a computational device used to define an order on vertices. Two equivalent rooted trees are distinct as ordered trees if the orders of their degree-3 vertices are different.

When ordering a rooted tree we need to take into account the hierarchical relationships between species which naturally determines a partial order on vertices as the ancestors must be ordered so that they existed before their children. Thus, in the absence of any time-related knowledge for speciation events, it seems proper to consider the vertex ordering as being generated uniformly at random from the orderings consistent with the ancestor-descendant relationships already defined in the species tree S, and this defines our prior on order. As an example, Figure 4.2 shows a rooted species tree S representing the phylogeny of seven species, five internal vertices and the root. The other two trees shown in Figure 4.2 are two possible ordered representations of S showing each vertex pseudotime.

Let $I(a, b)$ be the number of internal vertices on an ordered rooted tree, where a and b are vertices in the tree, whose associated pseudotimes lie strictly between $t(a)$ and $t(b)$. $I(a, b)$ counts the number of *intermediate* vertices between $t(a)$ and

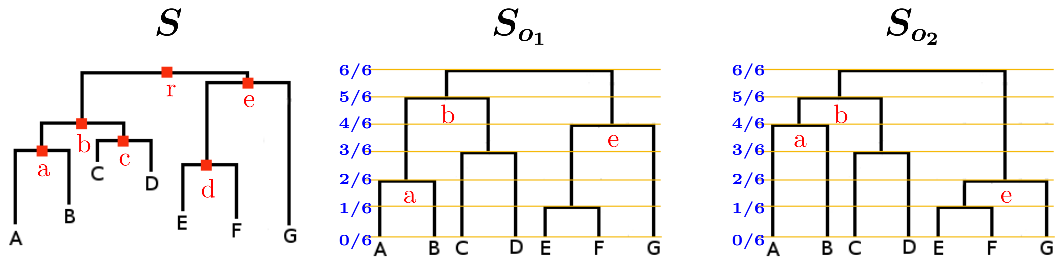


Figure 4.2: **Species tree S describes the phylogeny of seven extant species: A, B, C, D, E, F and G.** This tree also includes five internal vertices a, b, c, d, e and the root r. S_{o_1} and S_{o_2} represent two possible ordered species trees originated from S. Pseudotimes are shown on the left in blue. In S_{o_1}, vertex $t(a) = 2/6$ and $t(e) = 4/6$ while in S_{o_2} their pseudotimes are the opposite. In the prior for ordering, we consider the vertex ordering as being assigned uniformly at random, subject to the ancestry constraints on the tree.

$t(b)$. Taking the example of the two ordered rooted trees in Figure 4.2, for S_{o_1}, $I(a,b) = 2$ while S_{o_2} has $I(a,b) = 0$. We will also assume that the descendant vertices of a specific vertex u, such that $d(u) = 3$, can be divided in two subsets and that $s_L(u)$ (resp. $s_R(u)$) corresponds to the number of degree-3 vertices which are in the subtree on the left (resp. right) of u. In S_{o_1}, $s_L(b) = s_R(b) = 1$ while $s_L(e) = 1$ and $s_R(e) = 0$. Song (2006) has shown that the number of inequivalent ordered trees with n leaves, for $n \geq 2$, originated from all possible rooted trees with n leaves, is

$$\prod\nolimits_{k=2}^{n} \binom{k}{2} = \frac{n!(n-1)!}{2^{n-1}}. \tag{4.1}$$

Nevertheless, the number of inequivalent ordered trees resulting from a single rooted tree T (denoted plain by Song (2006)), $N_{ordered}(T)$, depends on T's topology and can be determined as:

$$N_{ordered}(T) = \prod_{i=\text{vertices in T of degree 2 or higher}} \mathit{\Delta}(i), \tag{4.2}$$

where

$$\Delta(i) = \frac{(s_L(i) + s_R(i))!}{s_L(i)! + s_R(i)!}. \tag{4.3}$$

Note that this is invariant under interchange of $s_L(u)$ and $s_R(u)$. See Song (2006) for more details. In terms of the hierarchical structure of the model, assuming an order on S results in the introduction of an extra level over Suchard's model (Figure 4.3).

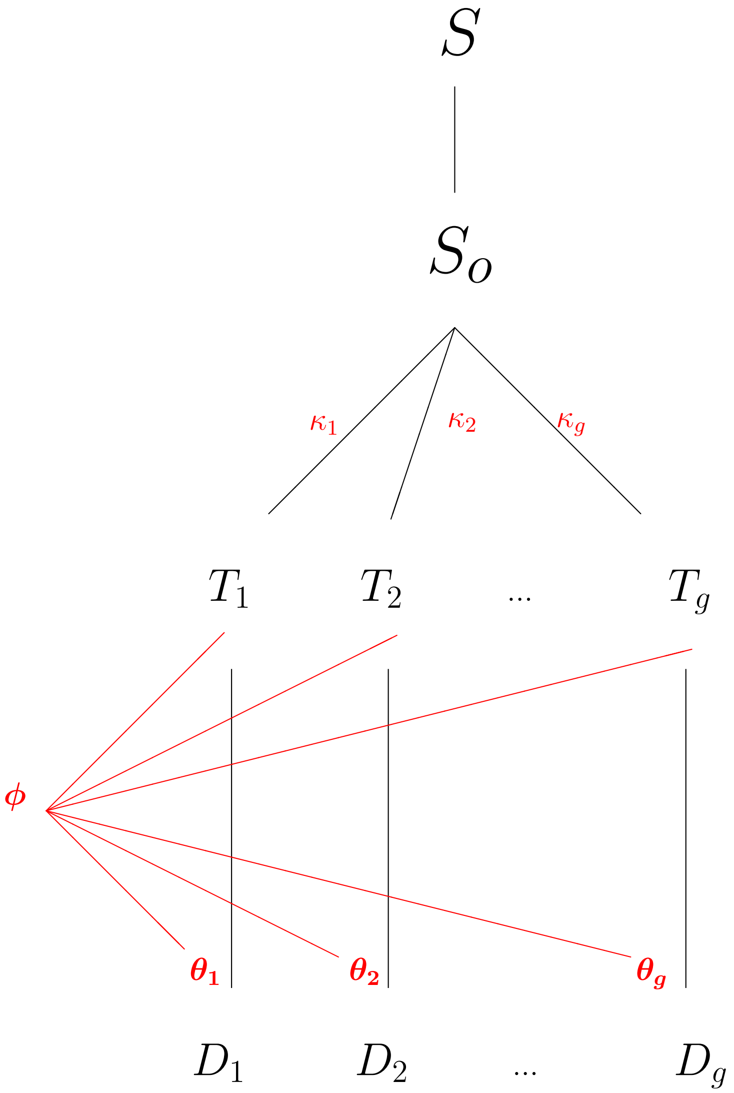


Figure 4.3: **Introducing a new level in Suchard's hierarchical phylogenetic model by imposing an order on S.** S_o represents the ordered species tree.

4.3 Species Contemporaneity on LGT events

Following Suchard's approach for modelling LGTs (see Section 3.3 for more details) we start by describing the collection of trees for n extant taxa as a mathematical graph and further assume a random walk on this graph that mirrors the observed effects of LGT.

Let $\mathscr{G} = (\Upsilon, \varepsilon)$ be a graph with vertex set Υ and edge set ε, where each vertex represents a tree on taxa $1, \ldots, n$ and an edge $uv \in \varepsilon$ corresponds to a certain relationship between the trees represented by adjacent vertices, $u, v \in \Upsilon$, on $\mathscr{G}$. The set of all vertices adjacent to a specific vertex are called its *neighbourhood* and several operators exist which can be used to specify the neighbourhood of a vertex and hence the structure of $\mathscr{G}$, the most common being NNI, SPR and TBR (Figure 2.4). The SPR operator offers an advantage over the other operators due to its potential biological interpretation as it mirrors the effect of an LGT on a phylogenetic tree. Consequently, the application of the SPR operator to a topology should represent the differences observed between a species tree and an individual gene tree affected by one LGT.

One important point is that the precise definition of the SPR operation depends on the type of tree on which the operation is being performed. The more constraints a tree has, the more restrictive an SPR operation has to be. For example, in an unrooted tree the operator simply selects and cuts any branch in the initial tree, pruning a subtree, and then regrafts this subtree by selecting and subdividing a preexisting branch in the remaining tree (see Figure 2.4(b)). The vertex of degree two that remained where the pruned edge used to be connected is deleted and the two remaining edges are replaced by a single edge in an operation called *forced contraction*, in order to maintain the binary property of the resulting tree. But when performing an SPR operation in unordered and ordered rooted trees several aspects need to be taken into consideration and different approaches need to be taken.

Song (2006) defines three different kinds of SPR operations on unordered rooted trees (from now on named as *rooted trees*), which are illustrated in Figure 4.4. Exemplifying the first type of SPR operation, tree T_1 was originated by cutting edge e_b, which is not connected to the root r, and regrafting it onto a preexisting edge, e_a, in the remaining part of T. In the second type of operation, because the pruned edge e_c is connected to r, when regrafted to e_a, the ancestor of s_1, s_2 and s_3 will become the root r' of T_2, while r and the edge connecting it to r' are eliminated. The third kind of operation allows the possibility of cutting an edge not connected with r (in the example of T_3 we cut edge e_b) and then creating a new root r'' and an edge connecting r and r''. Then the pruned subtree is joined to r''.

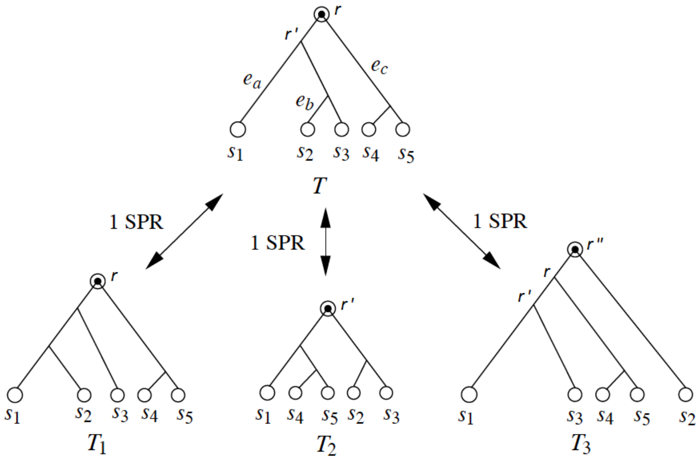


Figure 4.4: **Illustration of SPR operations on rooted trees.** (Figure reproduced from Song (2006).)

Nevertheless, in Song & Hein (2003), when studying the SPR operator on rooted phylogenetic trees, the authors reached the conclusion that, to determine correctly the minimum number of recombination events between two trees, the right kind of trees and the right kind of topological operation should be used, more specifically, they should be leaf-labelled rooted binary trees with totally ordered internal vertices. Therefore, and in order to perform SPR operations on ordered rooted trees, Song (2006) proposed an additional restriction on the definition of SPR operations. Let T be an ordered rooted tree. The SPR operation on T must satisfy the condition that, for any two vertices $v_i, v_j \in T$, if $t(v_i) < t(v_j)$ before the SPR operation, then $t(v_i) < t(v_j)$ after the SPR operation, and vice versa.

Song illustrates this situation with Figure 4.5 where it is shown that, if trees T_1 and T_2 are unordered rooted trees, then T_2 can be originated by only 1 SPR on T_1 by pruning the subtree containing l_4 and l_5 and regrafting it onto edge e. The same operation is not possible if T_1 and T_2 are ordered rooted trees as $t(v_2) > t(v_3)$ in T_1 but the opposite in T_2. Therefore, at least two SPR operations are required to transform T_1 into T_2.

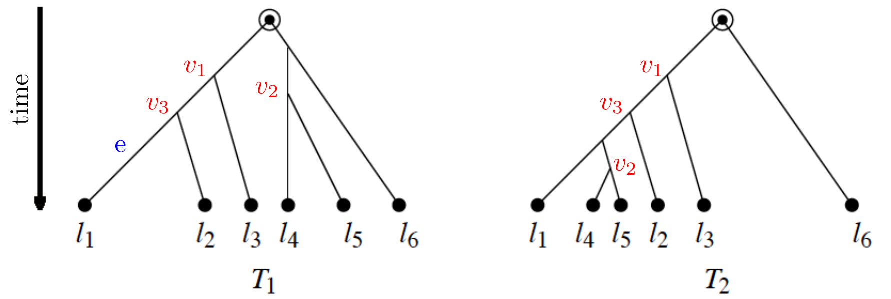


Figure 4.5: An example of trees which are more than one SPR operation apart if ordered yet only one SPR operation apart if unordered (figure adapted from Song (2006)).

From a biological point-of-view, although it seems to mirror the LGT effect, this operator has some drawbacks on unrooted, unordered and ordered rooted trees. Let us assume that any edge of the tree represents the amount of change that occurs between species u and its parent $p(u)$, and that a point on the edge represents a species v which, although represents the same species as u, it is genetically distinct as a result of the change that occurred over time.

Song's SPR on ordered rooted trees involves the deletion of all or part of the edge containing the prune location. Edges on an ordered rooted tree represent the existence of some species over some historical time interval. Deletion of edges, as with Song's SPR, changes the number of extant species at certain points in time. We need a properly defined SPR operation which preserves the number of extant species over any time interval since LGT also preserves this. In particular, this constraint would maintain the possibility of LGT between ancestral species which might be ruled out if any edges are deleted. The xSPR operation we define below, in fact maintains a bijection between points on an ordered rooted tree with pseudotimes and the same tree after xSPR (Figure 4.6).

Given this, we are proposing an extended version of the SPR operator (xSPR) as a more appropriate biological representation of an LGT event on an ordered rooted tree.

4.3.1 Extended SPR (xSPR)

Let T be an ordered rooted tree and let e_g and e_p be edges on T, which are contemporary, i.e., contain a shared pseudotime interval. The xSPR between e_g and e_p is defined as follows. The pruned edge e_p is cut at some point of its length, maintaining a *stumpy* edge available for further SPR moves. A vertex v_p is positioned at the free end of the stumpy edge and will have the same pseudotime as the new vertex v_g introduced in the graft edge e_g when attaching the pruned subtree. The vertex on the stumpy edge will represent the species that existed just before the occurrence of the LGT while v_g is the species that donated the gene. The binary property of the tree is maintained and one more edge and two new vertices are added. For a more visual explanation see Figure 4.6.

As in Suchard's model (see Subsection 3.3), in our model LGT will be modelled via an unweighted random walk process in the tree space graph, but in this case the graph is derived from the xSPR operator. In particular, vertices on the graph represent stumpy trees and each step on the random walk increases the number of stumps by one. Let $Y_{i,1}, \ldots, Y_{i,k_i}$ be a sequence of k_i LGTs for gene i on S_o. This determines a sequence of stumpy tree topologies $\tau_{i,0}, \tau_{i,1} \ldots, \tau_{i,k_i}$ where $\tau_{i,0} = S_o$ and τ_{i,k_i} is the final topology τ_i. Each $Y_{i,j}$ corresponds to a set $\{e_{p_{ij}}, e_{g_{ij}}, t_{ij}\}$, for $j = 1, ..., k_i$, where t_{ij} is the pseudotime of the j^{th} LGT for gene i, $e_{p_{ij}}$ the prune edge and $e_{g_{ij}}$ the graft edge. It is important to note that the relationship between the LGT history $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,k_i})$ and τ_i is not exclusive. Different LGT histories might originate the same topology, τ_i. However, the sequence of stumpy tree topologies $\tau_{i,0}, \tau_{i,1}, \ldots, \tau_{i,k_i}$ contains exactly the same information as $\boldsymbol{Y_i}$.

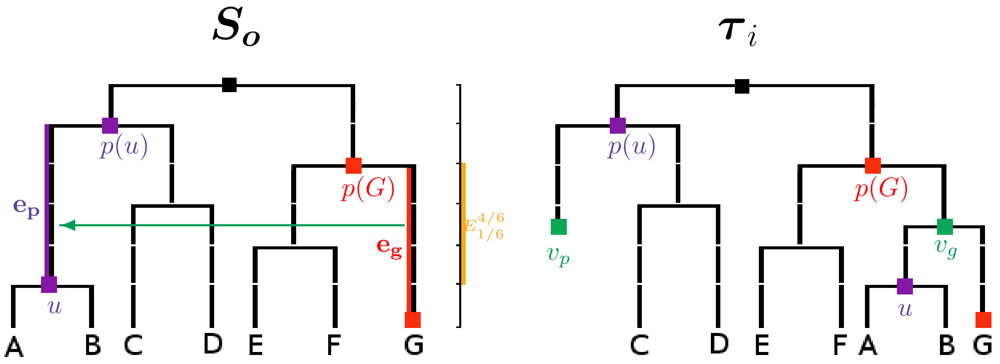


Figure 4.6: Generating a gene tree topology τ_i for gene i using a single xSPR operator.

Our prior on $\boldsymbol{Y_i}$, the location of xSPR events on the tree, takes the form

$$\pi(\mathbf{Y_i}|\kappa_i, S_o) = \pi(\tau_{i,1}|\tau_{i,0})\pi(\tau_{i,2}|\tau_{i,1}) \ldots \pi(\tau_{i\kappa_i}|\tau_{i(\kappa_i-1)}) \tag{4.4}$$

$$= \prod_{j=1}^{\kappa_i} \pi(\tau_{i,j}|\tau_{i,j-1}) \tag{4.5}$$

where each stumpy tree $\tau_{i,s}$ is conditional on the previous topology in the sequence, $\tau_{i,s-1}$.

Given $\tau_{i,j-1}$, the location of the next xSPR has the following distribution:

- The prune edge $e_{p_{ij}}$ is distributed uniformly at random from the edges on $\tau_{i,j-1}$.
- The graft edge $e_{g_{ij}}$ is distributed uniformly at random from the edges that are contemporary to $e_{p_{ij}}$.
- If the pseudotime interval shared by e_p and e_g contains one or more intermediate vertices $I(u, p(G)) > 1$, defining two or more pseudotime epochs within the interval, the epoch where the LGT occurs is distributed uniformly at random.
- Within the epoch, the specific pseudotime of an LGT is also distributed uniformly at random.

4.4 Model Specification

In the previous sections we described in detail the novel adaptations we have introduced to Suchard's Bayesian hierarchical model. In this section we provide a complete specification of our model. For an easier understanding, Figure 4.7 shows our hierarchical model framework and the associated parameters to each level, while the diagram in Figure 4.8 depicts the dependencies between parameters. In a more general model we might attempt to infer S jointly with the other parameters, but for the purpose of this thesis we consider S as known.

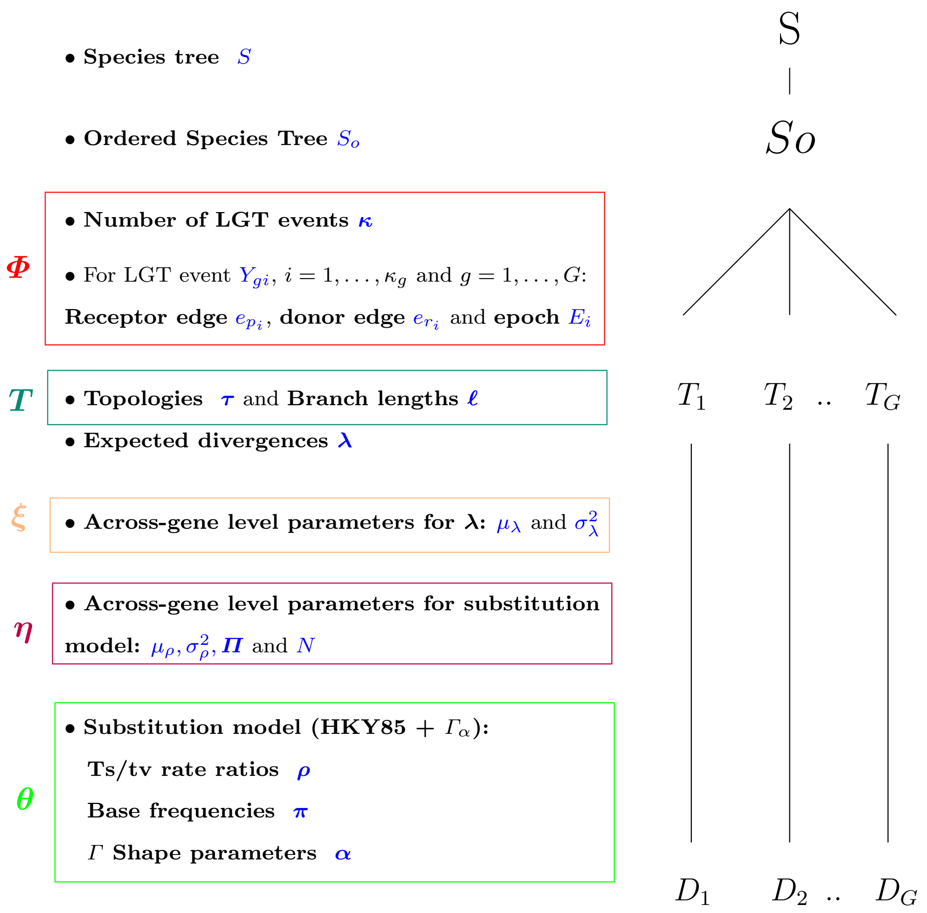

• Species tree S
• Ordered Species Tree S_o
$\boldsymbol{\Phi}$
• Number of LGT events $\boldsymbol{\kappa}$
• For LGT event Y_{gi}, $i = 1, \ldots, \kappa_g$ and $g = 1, \ldots, G$:
Receptor edge e_{p_i}, donor edge e_{r_i} and epoch E_i
$\boldsymbol{T}$
• Topologies $\boldsymbol{\tau}$ and Branch lengths $\boldsymbol{\ell}$
• Expected divergences $\boldsymbol{\lambda}$
$\boldsymbol{\xi}$
• Across-gene level parameters for $\boldsymbol{\lambda}$: μ_λ and σ^2_λ
$\boldsymbol{\eta}$
• Across-gene level parameters for substitution model: $\mu_\rho, \sigma^2_\rho, \boldsymbol{\Pi}$ and N
$\boldsymbol{\theta}$
• Substitution model (HKY85 + Γ_α):
Ts/tv rate ratios $\boldsymbol{\rho}$
Base frequencies $\boldsymbol{\pi}$
Γ Shape parameters $\boldsymbol{\alpha}$
S
So
T_1 T_2 .. T_G
D_1 D_2 .. D_G

Figure 4.7: Diagram depicting our novel Bayesian hierarchical phylogenetic model.

To facilitate the model's formal description our parameters will be grouped together in the following way:

- $\boldsymbol{\Phi} = (\boldsymbol{\Phi_1}, \ldots, \boldsymbol{\Phi_G})$ with $\boldsymbol{\Phi_g} = (\kappa_g, \mathbf{Y_g})$, for $g = 1, \ldots, G$, includes all LGT related parameters;

- $\boldsymbol{T} = (\boldsymbol{\tau}, \boldsymbol{\ell})$ denotes the set of gene trees;

- $\boldsymbol{\xi} = (\mu_\lambda, \sigma^2_\lambda)$ corresponds to the across-gene level parameters for gene divergence $\boldsymbol{\lambda}$;

- $\boldsymbol{\eta} = (\mu_\rho, \sigma^2_\rho, \boldsymbol{\Pi}, N)$ includes the substitution model across-gene level parameters;

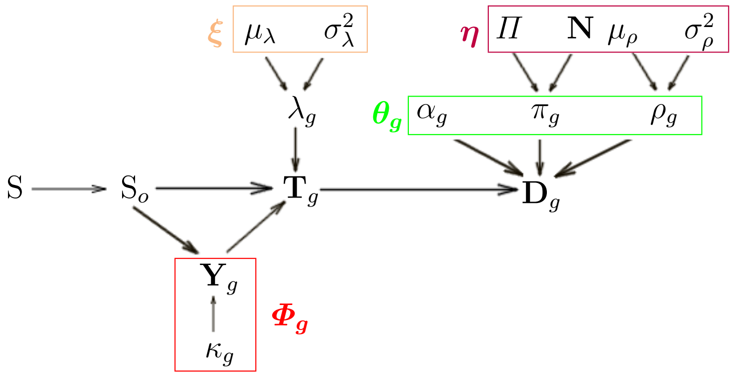


Figure 4.8: **Diagram depicting the parameter dependencies.** The parameters for a single gene tree T_g and corresponding alignment D_g are shown for simplicity. The full diagram would have these elements repeated G times.

- $\boldsymbol{\theta} = (\boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{\alpha})$ corresponds to all substitution model parameters.

Therefore, the set $(S_o,\boldsymbol{\Phi},\boldsymbol{T},\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{\theta})$ specifies the complete model and the model's joint posterior distribution can be written as

$$\pi(S_o, \boldsymbol{\Phi}, \boldsymbol{T}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{D}|S) = \pi(S_o|S)\pi(\boldsymbol{\xi})\pi(\boldsymbol{\eta}) \tag{4.6}$$
$$\times \prod_{g=1}^{G} \pi(\tau_g|\boldsymbol{Y_g}, S_o)\pi(\boldsymbol{Y_g}|\kappa_g, S_o)\pi(\kappa_g)\pi(\boldsymbol{\ell}_g|\lambda_g)\pi(\lambda_g|\boldsymbol{\xi})\pi(\boldsymbol{\theta}_g|\boldsymbol{\eta})\pi(D_g|\boldsymbol{\theta}_g, \tau_g, \boldsymbol{\ell}_g).$$

4.4.1 Priors

The aim in this section is to define each term in Equation 4.6 explicitly.

Prior on S_o

The distribution $\pi(S_o|S)$ is uniform on all orders o compatible with the ancestry relations in S, as stated in Section 4.2.

Priors on $\boldsymbol{\Phi}$ and $\boldsymbol{T}$

As the number of LGT events κ_g, the LGT history $\boldsymbol{Y_g}$ and gene trees $\boldsymbol{T_g}$ are dependent, there is a need to handle them jointly. Their priors are as follows:

Number of LGTs. We use truncated Geometric priors for the number of LGTs, specifically $\kappa_g \sim TGeom(a_\kappa, b_\kappa)$, for $a_\kappa = 0.5$ and $b_\kappa = B/2$, where B is the number of edges on S. The need for a truncation on the maximum number of LGTs allowed resides in the fact that, for a high number of LGTs, the level of perturbation on the tree might be such that any relation with the initial tree is lost. As the expected number of LGTs between species is generally low, we assumed that the maximum number of LGTs is half the number of edges.

LGT histories. For each LGT event $\boldsymbol{Y_{gi}} = (e_{p_{gi}}, e_{g_{gi}}, E_{gi})$, the prune edge $e_{p_{gi}}$, graft edge $e_{g_{gi}}$ and epoch E_{gi} are chosen uniformly at random from the set of viable choices. These viable choices must respect the species contemporaneity, as described in Subsection 4.3.1. Each topology τ_g, for $g = 1, \ldots, G$, is a deterministic function of the set of LGT events $\boldsymbol{Y}_g = (\boldsymbol{Y}_{g,1}, \ldots, \boldsymbol{Y}_{g,\kappa_g})$ on S_o, i.e, $\pi(\tau_j|\boldsymbol{Y}_g, S_o)$ is an indicator function.

Branch lengths. A common approach in Bayesian phylogenetics is to use exponential priors on branch lengths. We assume a hierarchical prior for the branch lengths such that

$$\ell_{gj}|\lambda_g \overset{indep}{\sim} SExp(\lambda_g, x), \qquad j = 1, \ldots, B \tag{4.7}$$

$$\lambda_g|\mu_\lambda, \sigma^2_\lambda \sim LN(\mu_\lambda, \sigma^2_\lambda) \tag{4.8}$$

$$\mu_\lambda \sim N(m_\lambda, v_\lambda) \tag{4.9}$$

$$\frac{1}{\sigma^2_\lambda} \sim Gamma(a_\lambda, b_\lambda), \tag{4.10}$$

where $SExp(\lambda_g, x)$ defines a shifted exponential distribution taking values of at least $x = 0.002$ and has rate λ_g. Also, B is the number of branch lengths. The expected divergence λ_g is unknown and has lognormal prior with across-gene level unknown mean μ_λ and unknown variance σ^2_λ. As in Suchard's approach, to allow direct Gibbs sampling of (μ_λ, σ^2_λ) we give them conjugate priors (see Subsection 2.2.2 for more details on conjugate priors). We assume $m_\lambda = 0$, $v_\lambda = 10$, $a_\lambda = 2.1$ and $b_\lambda = 1.1$. These settings result in relatively uninformative priors for both parameters, and sets the prior expectation and variance for σ_λ to 0 and 10, respectively. Equations (4.7)-(4.10) determine the prior component distributions $\pi(\boldsymbol{\ell_g}|\lambda_g)$, $\pi(\lambda_g|\boldsymbol{\xi})$ and $\pi(\boldsymbol{\xi})$.

This hierarchical structure allows the information sharing between the genes as discussed before and we further assume that branch lengths cannot be smaller than 0.002. The prior assigns probability 0 to branch lengths smaller than 0.002, as a branch length of 0.002 represents the occurrence of 1 substitution every 500 nucleotides on average. In our simulation studies presented in Chapter 5, where the simulated sequences were 1000 nucleotides long, we were unable to infer very small branch lengths, as the Markov chains tended to 0, leading to the conclusion that the data might not have enough information to infer very small branch lengths.

Priors on $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$

We will discuss next the priors for each specific parameter in $\boldsymbol{\theta}$.

Transition:transversion rate ratio $\boldsymbol{\rho}$, $\boldsymbol{\mu_\rho}$ and $\boldsymbol{\sigma^2_\rho}$

Similarly to λ_g, a hierarchical prior will be assumed for ρ_g such that

$$\rho_g|\mu_\rho, \sigma^2_\rho \sim LN(\mu_\rho, \sigma^2_\rho) \tag{4.11}$$

$$\mu_\rho \sim N(m_\rho, v_\rho) \tag{4.12}$$

$$\frac{1}{\sigma^2_\rho} \sim Gamma(a_\rho, b_\rho). \tag{4.13}$$

The last two equations define semi-conjugate priors, and we take $m_\rho = 0$, $v_\rho = 10$, $a_\rho = 2.1$ and $b_\rho = 1.1$ as in (Suchard *et al.*, 2003).

Base frequencies $\boldsymbol{\pi}$, $\boldsymbol{\Pi}$ and $\boldsymbol{N}$

The distributions $\boldsymbol{\pi}_g$ are defined on the simplex $\mathcal{S}^3 \subset \mathbb{R}^4$ and are naturally modelled by a Dirichlet distribution with

$$\boldsymbol{\pi}_g|\boldsymbol{\Pi}, N \sim Dir(\boldsymbol{\Pi}N), \tag{4.14}$$

where $\boldsymbol{\Pi} = (\Pi_A, \Pi_G, \Pi_C, \Pi_T)$ are the across-gene level proportions for each nucleotide, and N is a pseudocount measure of precision across $\boldsymbol{\pi}$. Again, a hierarchical prior is assumed on $\boldsymbol{\pi}$ by taking

$$\boldsymbol{\Pi} \sim Dir(M), \tag{4.15}$$

for $M = (1, 1, 1, 1)$, which provides a flat prior on $\boldsymbol{\Pi}$, and

$$N \sim Gamma(a_N, b_N), \tag{4.16}$$

for $a_N = b_N = 0.1$. These are the choices of prior parameters taken by Suchard *et al.* (2003).

Gamma shape parameter $\boldsymbol{\alpha}$

The value of α_g is modelled using an exponential prior, so that the MCMC procedure can explore different shapes of the gamma distribution associated with the evolutionary rate among sites. Therefore the shape parameter α_g for each gene is given by an independent Exponential prior

$$\alpha_g \overset{indep}{\sim} Exp(a_\alpha), \tag{4.17}$$

with $a_\alpha = 1$, which defines a distribution with mean and variance equal to 1. This centres the distribution of the site heterogeneity rates r_i to be exactly between an L-shape distribution ($\alpha_g \leq 1$) and a more bell-shape distribution ($\alpha_g > 1$).

4.4.2 Updating the parameters

The majority of the parameters are updated in single moves, with all other parameters fixed, with the exception of the parameters related to the LGT history and gene tree topologies, which are updated jointly. Next we will describe the proposal mechanisms for each parameter as well as the respective acceptance probabilities.

Novel MCMC proposal for the order of $\boldsymbol{S_o}$

When proposing a new order o, it is important that the current set of LGT histories, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_g$, is compatible with the proposed order, as otherwise we would need a joint proposal on o and $\boldsymbol{Y}$. Thus, the proposal distribution for S_o should be conditional not only on S but also on the corresponding LGT histories for the current gene trees. Any proposed ordering needs to allow the occurrence of the current LGT histories for all genes, and not rule them out.

The novel proposal mechanism for S_o that we have developed will propose a change in the order of one vertex v at each iteration. The vertex will be chosen uniformly at random, such that

$$v \sim U(V) \tag{4.18}$$

where $V = \{v_1, v_2, \ldots, v_{n-2}\}$ is the set of degree-3 vertices of S_o. The new order will be achieved by assigning a new pseudotime to v. This new time must not:

- Contradict ancestor-descendant ordering on S;
- Invalidate any LGT event relating the gene trees and S, by moving edges involved in the corresponding SPR and making them non-contemporary.

Given these assumptions a new pseudotime for v will be proposed by considering the following:

1. Vertex v will be allowed to move only to epochs between pseudotimes t_{min} and t_{max} defined by

$$\begin{aligned} t_{min} = \max\{&\max\{t(u) : u \text{ is a child of } v\},\\ &\max_{g,i}\{t_i : (t_i, e_{p_i}, e_{g_i}) \in \boldsymbol{Y_g} \text{ and } e_{p_i} \text{ or } e_{g_i} \text{ descend from } v\}\} \end{aligned}$$

$$t_{max} = \min\{\min\{t(u) : u \text{ is the ancestor of } v\},$$
$$\min_{g,i}\{t_i : (t_i, e_{p_i}, e_{g_i}) \in \boldsymbol{Y_g} \text{ and } e_{p_i} \text{ or } e_{g_i} \text{ is the ancestor edge of } v\}\}$$

This ensures we satisfy the assumptions above.

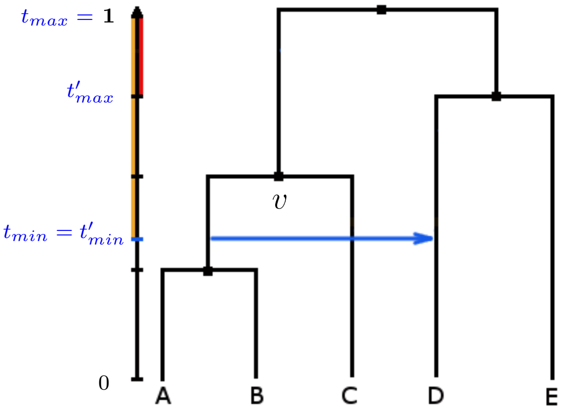


Figure 4.9: **Diagram depicting an example of defining the pseudotime interval to propose a new pseudotime for vertex v.** The blue arrow represents an LGT from a descendent of v and an ancestral species of D. Time is defined between 0 and 1, where 0 corresponds to present time and 1 the time defined by the ancestor of all species in the study. Although v is allowed to move to any time point on the interval signalled in orange, its order will change only when moving to the epoch $[t'_{max}, t_{max}]$. In this example $[t'_{min}, t_{min}] = \emptyset$.

2. Certain changes to the pseudotime of v, more specifically, moving v to any of its adjacent epochs, will result in no change to the order of vertices, and so we want to rule these out. Therefore we will define

$$t'_{min} = \max\{\max_{u}\{t(u) : t(u) < t(v)\},$$
$$\max_{g,i}\{t_i : (t_i, e_{p_i}, e_{g_i}) \in \boldsymbol{Y_g} \text{ and } e_{p_i} \text{ or } e_{g_i} \text{ descend from } v \text{ and } t_i < t(v)\}\}$$

$t'_{max} = \min\{\min\{t(u) : t(u) > t(v)\},$

$\min_{g,i}\{t_i : (t_i, e_{p_i}, e_{g_i}) \in \boldsymbol{Y_g}$ and e_{p_i} or e_{g_i} is the ancestor edge of v and $t_i >$ $t(v)\}\}.$

Now we have two subintervals $[t_{min}, t'_{min}]$ and $[t_{max}, t'_{max}]$ where a vertex ordering will occur and pseudotime $t^*(v)$ will be uniformly at random chosen, with

$$t^*(v) \sim U([t_{min}, t'_{min}] \cup [t_{max}, t'_{max}]). \tag{4.19}$$

The proposal density is

$$q(S_o^*|S_o, \boldsymbol{Y}) = \frac{1}{(t_{max} - t'_{max}) + (t'_{min} - t_{min})} \times \frac{1}{n-2} \tag{4.20}$$

where $n-2$ is the number of internal vertices on S_o. See Figure 4.9 for an explanatory diagram.

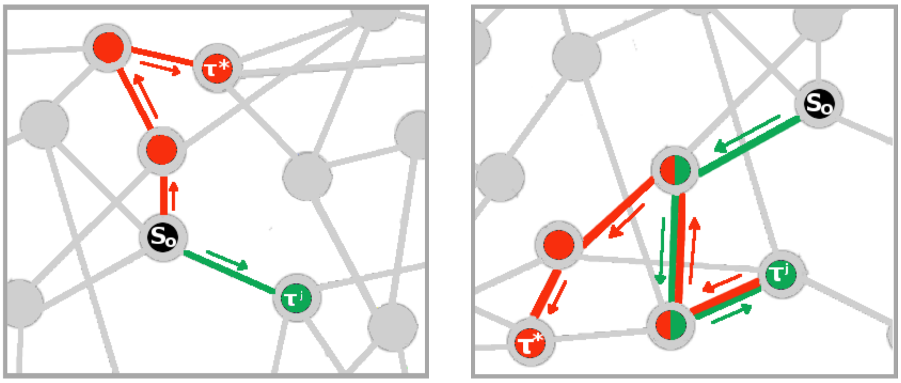
Novel MCMC proposals for the Number of LGTs, LGT history and Gene Trees

A gene tree T_g for gene g comprises a topology τ_g resultant from κ_g xSPRs operations in tree space, where $\boldsymbol{Y_g} = (Y_{g,1}, \ldots, Y_{g,\kappa_g})$ is the ordered set of the xSPRs which originated τ_g from S_o, and a set of B branch lengths, $\boldsymbol{\ell_g} = (\ell_{g,1}, \ldots, \ell_{g,B})$, each one assigned to a branch of τ_g. The sequence of xSPRs is associated with a sequence of stumpy tree topologies $(\tau_{g,0}, \ldots, \tau_{g,\kappa_g})$, where $\tau_{g,0} = S_o$ and $\tau_{g,\kappa_g} = \tau_g$. These stumpy tree topologies are a deterministic function of $\boldsymbol{Y_g}$ and S_o, and represent equivalent information to $\boldsymbol{Y_g}$. Therefore, it is sensible to jointly propose κ_g^*, $\boldsymbol{Y}_g^*$, τ_g^*, and $\boldsymbol{\ell_g^*}$. Three different proposal mechanisms were implemented assuming fixed S and S_o.

1. **Independence proposal**

   A simple approach is an independence sampler where for gene g and iteration j of the MCMC sampler, we will generate κ_g^* from the prior and perform κ_g^* xSPR moves on S_o generating $\boldsymbol{Y}_g^*$ again by sampling from the prior (see Figure 4.10), and hence τ_g^*. Conditional on τ_g^* and λ_g, the branch lengths are sampled from the prior too. The proposal distribution is

$$q(\boldsymbol{\ell}_g^*, \boldsymbol{Y}_g^*, \kappa_g^* | S_o, \boldsymbol{Y}_g, \kappa_g) = \pi(\kappa_g^*)\pi(\boldsymbol{Y}_g^* | S_o, \kappa_g^*)\pi(\boldsymbol{\ell}_g^* | \lambda_g). \tag{4.21}$$



(a) Independence sampler.

(b) Backward-forward sampler.

Figure 4.10: **Topology proposals:** a) The current state topology is τ^j. For every proposal, the random walk starts on S_o and for κ^* xSPR moves (in this case $\kappa^* = 3$, in red), a new topology τ^* will be proposed in iteration $j+1$. This topology is independent from τ^j given S_o; b) Assuming τ^j is $\kappa^* = 3$ xSPRs away from S_o (in green), in this example, a new topology τ^* will be proposed using $b^* = 2$ xSPRs backward and $f^* = 2$ xSPRs forward (in red).

2. **Backward-forward proposal**

   A more sophisticated approach, able to propose τ_g^* using the information in τ_g is the *backward-forward sampler*. The idea is to propose $\boldsymbol{Y}_g^*$ by taking $\boldsymbol{Y}_g$, ignoring the last b^* xSPR in $\boldsymbol{Y}_g^*$, but then appending f^* new xSPRs on to the resultant xSPR history.

Let b^* be the proposed number of xSPRs ignored ('backward') in the current $\boldsymbol{Y}_g$ and f^* the number of steps appended ('forward') i.e. applied to τ_{g,κ_g-b^*} (see Figure 4.10). In this case

$$\kappa_g^* = \kappa_g - b^* + f^* \tag{4.22}$$

We will propose b^* and f^* in the following way:

- If $\kappa_g = 0$, the current topology $\tau_g = S_o$ and no back steps can be taken, i.e. $b^* = 0$. Thus, we will propose only steps forward such that $f^* \sim Po(d)$, for $d = 2$.
- For $\kappa_g \neq 0$, $b^* \sim TPo(e, \kappa_g)$, for $e = 1$. The proposal for f^* then depends on the value of b^*.
  (a) If $b^* = 0$ then $f^* \sim Po(h)$, for $h = 0.5$.
  (b) Else, $f^* \sim Po(b^*)$.

This determines $q(f^*, b^*|\kappa_g)$. Given b^* and f^*, the f^* xSPRs 'forward' from τ_{g,κ_g-b^*} are proposed from the prior. In order to compute the proposal ratio it is convenient in this section to drop the subscript g and write τ_j instead of $\tau_{g,j}$ for a stumpy tree topology in $\boldsymbol{\tau}_g = (\tau_{g,0}, \tau_{g,0}, \ldots, \tau_{g,k_g})$, and similarly for κ_g and $\boldsymbol{Y}_g$. Furthermore, the reverse move from $\boldsymbol{Y}^*$ to $\boldsymbol{Y}$ has

- $b = f^*$ steps back, and
- $f = b^*$ steps forward.

In this notation, the proposed sequence of stumpy tree topologies corresponding to $\boldsymbol{Y}^*$ is $\tau_0, \tau_1, \ldots, \tau_{\kappa-b^*}, \tau^*_{\kappa-b^*+1}, \ldots, \tau^*_{\kappa-b^*+f^*}$. The proposal density is

$$\begin{aligned} q(\boldsymbol{Y}^*, \kappa^* | \boldsymbol{Y}, \kappa, S_o) &= q(f^*, b^* | \kappa) q(\boldsymbol{Y}^* | \boldsymbol{Y}, f^*, b^*, \kappa, S_o) \\ &= q(f^*, b^* | \kappa) \pi(\tau^*_{\kappa-b^*+1} | \tau_{\kappa-b^*}) \times \pi(\tau^*_{\kappa-b^*+2} | \tau^*_{\kappa-b^*+1}) \\ &\qquad \times \ldots \times \pi(\tau^*_{\kappa-b^*+f^*} | \tau^*_{\kappa-b*+f^*-1}) \\ &= q(f^*, b^* | \kappa) \pi(\tau^*_{\kappa-b^*+1} | \tau_{\kappa-b^*}) \times \left(\prod_{j=\kappa-b^*+2}^{\kappa-b^*+f^*} \pi(\tau^*_j | \tau^*_{j-1}) \right). \end{aligned} \tag{4.23}$$

Each term of the form $\pi(\tau_+, \tau_-)$, where τ_+ and τ_- are stumpy tree topologies related by a single xSPR from τ_- to τ_+, corresponds to the prior on xSPRs described in Section 4.3.1. In order to compute the acceptance probability later, we need the following calculation of a quantity denoted A_{top} (which is the part of the acceptance probability attributed to the topology):

$$\begin{aligned} A_{top} &= \frac{\pi(\kappa^*)}{\pi(\kappa)} \frac{\pi(\boldsymbol{Y}^* | \kappa^*, S_o)}{\pi(\boldsymbol{Y} | \kappa, S_o)} \frac{q(\boldsymbol{Y}, \kappa | \boldsymbol{Y}^*, \kappa^*, S_o)}{q(\boldsymbol{Y}^*, \kappa^* | \boldsymbol{Y}, \kappa, S_o)} \\ &= \frac{\pi(\kappa^*)}{\pi(\kappa)} \frac{\pi(\boldsymbol{Y}^* | \kappa^*, S_o)}{\pi(\boldsymbol{Y} | \kappa, S_o)} \times \frac{q(b, f | \kappa^*)}{q(b^*, f^* | \kappa)} \frac{q(\boldsymbol{Y} | \boldsymbol{Y}^*, f^*, b^*, \kappa^*, S_o)}{q(\boldsymbol{Y}^* | \boldsymbol{Y}, f, b, \kappa, S_o)} \\ &= \frac{\pi(\kappa^*)}{\pi(\kappa)} \frac{q(b, f | \kappa^*)}{q(b^*, f^* | \kappa)} \\ &\times \frac{\left(\prod_{j=1}^{\kappa-b^*} \pi(\tau_j | \tau_{j-1}) \right) \times \pi(\tau^*_{\kappa-b^*+1} | \tau_{\kappa-b^*}) \times \left(\prod_{j=\kappa-b^*+2}^{\kappa-b^*+f^*} \pi(\tau^*_j | \tau^*_{j-1}) \right)}{\prod_{j=1}^{\kappa} \pi(\tau_j | \tau_{j-1})} \end{aligned} \tag{4.24}$$

$$\times \frac{\pi(\tau_{\kappa-b^*+1} | \tau_{\kappa-b^*}) \times \left(\prod_{j=\kappa-b^*+2}^{\kappa} \pi(\tau_j | \tau_{j-1}) \right)}{\pi(\tau^*_{\kappa-b^*+1} | \tau_{\kappa-b^*}) \times \left(\prod_{j=\kappa-b^*+2}^{\kappa-b^*+f^*} \pi(\tau^*_j | \tau^*_{j-1}) \right)} \tag{4.25}$$

$$= \frac{\pi(\kappa^*)}{\pi(\kappa)} \frac{q(b, f | \kappa^*)}{q(b^*, f^* | \kappa)},$$

as all the other terms cancel. Equations 4.24 and 4.25 correspond to the prior and proposal ratios, respectively.

Conditional on τ^* (the proposed gene tree topology), branch lengths for τ^* are proposed independently from the prior, that is

$$\ell_j^* \overset{indep}{\sim} Exp(\lambda), \quad j = 1, \ldots, B. \tag{4.26}$$

This determines $q(\boldsymbol{\ell}^*|\tau^*, \lambda)$. The acceptance probability for the full proposal on topology and edges lengths is $\min\{1, A\}$ where

$$A = A_{top} \times \frac{\pi(\boldsymbol{\ell}^*|\lambda)}{\pi(\boldsymbol{\ell}|\lambda)} \times \frac{\pi(D|\theta, \tau^*)}{\pi(D|\theta, \tau)} \times \frac{q(\boldsymbol{\ell}|\tau, \lambda)}{q(\boldsymbol{\ell}^*|\tau^*, \lambda)} \tag{4.27}$$

As the branch lengths are proposed from the prior, the prior and proposal ratios will cancel leaving us with

$$A = \frac{\pi(\kappa^*)}{\pi(\kappa)} \frac{q(b, f|\kappa^*)}{q(b^*, f^*|\kappa)} \times \frac{\pi(D|\theta, \tau^*)}{\pi(D|\theta, \tau)} = A_{top} \times \frac{\pi(D|\theta, \tau^*)}{\pi(D|\theta, \tau)}. \tag{4.28}$$

3. **One-step backward-forward proposal**

The one-step backward-forward proposal is a special case of the backward-forward proposal, where it proposes moves only one xSPR forward, or one xSPR backward from the current state. When the chain's current state for topology is S_o then it only moves forward. These smaller local moves increases the acceptance probability by producing less extensive disturbances on the LGT history and resulting topology. In this situation, the proposal ratio of the proposed moves depends on κ_g and the maximum number of LGTs allowed by the model κ_{gM}. For $\kappa_g = 0$, the probability of proposing one step backward is 0, $\pi(b = 1) = 0$ since the current topology is S_o, and a step forward will be proposed with probability 1.0. If $0 < \kappa_g < \kappa_{gM}$, then $\pi(b = 1) = \pi(f = 1) =$

0.5. For $\kappa_g = \kappa_{gM}$, no steps forward are allowed, and therefore $\pi(f = 1) = 0$, while $\pi(b = 1) = 1$. The proposal ratios for every possible move are described in Table 4.1.

		Current κ_g						
		0	1	2	.	.	$\kappa_{g(M-1)}$	κ_{gM}
Proposed κ_g	0	-	2	-	.	.	-	-
	1	0.5	-	1	.	.	-	-
	2	-	1	-	.	.	-	-
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	$\kappa_{g(M-1)}$	-	-	-	.	.	1	0.5
	κ_{gM}	-	-	-	.	.	2	-

Table 4.1: Proposal ratios $q(\kappa_g|\kappa_g^*)/q(\kappa_g^*|\kappa_g)$ for the one-step backward-forward proposal.

For this proposal, and conditional on τ, only the branch lengths that changed during the topological part of the proposal are proposed from the prior, while all other branch lengths will be proposed independently from a log normal random walk centred on the current value, that is

$$\ell_j^* \overset{indep}{\sim} LN(\ell_j, v_\ell), \quad j = 1, .., B_u, \tag{4.29}$$

where B_u is the number of unchanged branch lengths and $v_\ell = 0.01$, which proved to provide good mixing. In this case, the prior and proposal ratios related to the unchanged branch lengths will contribute to the acceptance rate described in Equation 4.27, with $q(\boldsymbol{\ell}^*|\tau^*, \lambda) = \prod_{i=1}^{B_u} q(\ell_i^*|\tau^*, \lambda)$.

Update step for branch lengths $\boldsymbol{\ell}$

To provide proper mixing also for the branch lengths, with all other parameters fixed, new branch lengths are proposed from a log normal random walk centred on the current value. Two different proposals were developed:

- **Single branch length proposal.** Each branch length ℓ_{ij} is independently proposed from the proposal density defined in Equation 4.29 while all other branch lengths are fixed. This proposal, although time consuming, leads to a better acceptance rate.
- **All branch lengths proposal.** All branch lengths are proposed from the the same lognormal proposal density simultaneously as a joint proposal.

Update steps for $\boldsymbol{\lambda}$, $\boldsymbol{\mu_\lambda}$ and $\boldsymbol{\sigma^2_\lambda}$

In terms of proposals, new values for λ_g are proposed from a lognormal random walk such that

$$\lambda_g^*|\lambda_g \sim LN(\lambda_g, z_\lambda), \tag{4.30}$$

for $z_\lambda = 0.5$, which determines $q(\lambda_g^*|\lambda_g)$. As any changes on λ_g will not affect the likelihood calculation, the acceptance probability can be written as $\min(1, A)$ where

$$A = \frac{\pi(\lambda_g^*|\boldsymbol{\xi})}{\pi(\lambda_g|\boldsymbol{\xi})} \times \frac{q(\lambda_g|\lambda_g^*)}{q(\lambda_g^*|\lambda_g)} = \frac{\pi(\lambda_g^*|\boldsymbol{\xi})}{\pi(\lambda_g|\boldsymbol{\xi})} \times \frac{\lambda_g^*}{\lambda_g}. \tag{4.31}$$

In order to sample the across-gene level parameters μ_λ and σ^2_λ in the outer Metropolis-within-Gibbs cycle, we have derived their full conditional distributions to use Gibbs sampling. Therefore, we have that

$$\mu_\lambda|\lambda_g, \sigma^2_\lambda \sim N(\nu_{\mu_\lambda}, \sigma^2_{\mu_\lambda}) \tag{4.32}$$

and

$$\frac{1}{\sigma^2_\lambda}\bigg|\lambda_g, \mu_\lambda \sim Gamma\left(a_\lambda + \frac{G}{2}, b_\lambda + \frac{1}{2}SS_{\mu_\lambda}\right) \tag{4.33}$$

where $\nu_{\mu_\lambda} = \frac{m_\lambda \sigma_\lambda^2 + v_\lambda \sum_{g=1}^{G} \log \lambda_g}{\sigma_\lambda^2 + G v_\lambda}$, $\sigma^2_{\mu_\lambda} = \left(\frac{G}{\sigma_\lambda^2} + \frac{1}{v_\lambda}\right)^{-1}$ and $SS_{\mu_\lambda} = \sum_{g=1}^{G} (\mu_\lambda - \log \lambda_g)^2$.

Transition:transversion rate ratio $\boldsymbol{\rho}$, $\boldsymbol{\mu_\rho}$ and $\boldsymbol{\sigma^2_\rho}$

New values for ρ_g are proposed from a lognormal random walk such that

$$\rho_g^* | \rho_g \sim LN(\rho_g, z_\rho), \tag{4.34}$$

with $z_\rho = 0.1$, as this choice showed to provide good mixing. This determines $q(\rho_g^*|\rho_g)$. Therefore, the acceptance probability is $\min(1, A)$ with

$$\begin{aligned} A &= \frac{\pi(\rho_g^*|\mu_\rho, \sigma_\rho^2)}{\pi(\rho_g|\mu_\rho, \sigma_\rho^2)} \times \frac{\pi(D|\theta^*, \tau)}{\pi(D|\theta, \tau)} \times \frac{q(\rho_g|\rho_g^*)}{q(\rho_g^*|\rho_g)} &(4.35)\\ &= \frac{\pi(\rho_g^*|\mu_\rho, \sigma_\rho^2)}{\pi(\rho_g|\mu_\rho, \sigma_\rho^2)} \times \frac{\pi(D|\theta^*, \tau)}{\pi(D|\theta, \tau)} \times \frac{\rho^*}{\rho}. &(4.36)\end{aligned}$$

Across-gene level parameters μ_ρ and σ^2_ρ are sampled through Gibbs sampling as their full conditional distributions are

$$\mu_\rho | \rho_g, \sigma_\rho^2 \sim N(\nu_{\mu_\rho}, \sigma^2_{\mu_\rho}) \tag{4.37}$$

and

$$\frac{1}{\sigma_\rho^2}\bigg| \rho_g, \mu_\rho \sim Gamma\left(a_\rho + \frac{G}{2}, b_\rho + \frac{1}{2} SS_{\mu_\rho}\right), \tag{4.38}$$

where $\nu_{\mu_\rho} = \frac{m_\rho \sigma_\rho^2 + v_\rho \sum_{g=1}^{G} \log \rho_g}{\sigma_\rho^2 + G v_\rho}$, $\sigma^2_{\mu_\rho} = \left(\frac{G}{\sigma_\rho^2} + \frac{1}{v_\rho}\right)^{-1}$ and $SS_{\mu_\rho} = \sum_{g=1}^{G} (\mu_\rho - \log \rho_g)^2$.

Base frequencies ($\boldsymbol{\pi}$)

New values $\boldsymbol{\pi}_g^*$ will be proposed from a Dirichlet distribution centered on the current values $\boldsymbol{\pi}_g$ such that

$$\boldsymbol{\pi}_g^* \mid \boldsymbol{\pi}_g \sim Dir(n_\pi \boldsymbol{\pi}_g), \tag{4.39}$$

where $n_\pi = 500$, as this proved to provide good mixing. With this determining $q(\boldsymbol{\pi}_g^*|\boldsymbol{\pi}_g)$, the acceptance probability for this parameter is min $(1, A)$ where

$$A = \frac{\pi(\boldsymbol{\pi}_g^*|\boldsymbol{\Pi}, N)}{\pi(\boldsymbol{\pi}_g|\boldsymbol{\Pi}, N)} \times \frac{\pi(D|\theta^*, \tau)}{\pi(D|\theta, \tau)} \times \frac{q(\boldsymbol{\pi}_g|\boldsymbol{\pi}_g^*)}{q(\boldsymbol{\pi}_g^*|\boldsymbol{\pi}_g)}. \tag{4.40}$$

Across-gene level parameters are updated through Metropolis-Hastings steps such that

$$\boldsymbol{\Pi}^*|\boldsymbol{\Pi}, N, \boldsymbol{\pi} \sim Dir(n_\Pi \boldsymbol{\Pi}), \tag{4.41}$$

and

$$N^*|N, \boldsymbol{\Pi}, \boldsymbol{\pi} \sim LN(N, v_N), \tag{4.42}$$

where $n_\Pi = 300$ and $v_N = 0.1$. The acceptance probability in both cases does not depend on the likelihood and can be written as $\min(1, A)$ where

$$A = \frac{\pi(\boldsymbol{\Pi}^*)}{\pi(\boldsymbol{\Pi})} \times \frac{q(\boldsymbol{\Pi}|\boldsymbol{\Pi}^*)}{q(\boldsymbol{\Pi}^*|\boldsymbol{\Pi})}, \tag{4.43}$$

and

$$\begin{align} A &= \frac{\pi(N^*)}{\pi(N)} \times \frac{q(N|N^*)}{q(N^*|N)} \tag{4.44} \\ &= \left(\frac{N}{N^*}\right)^{1-a_N} \exp\left(b_N(N - N^*)\right) \times \frac{N^*}{N} \tag{4.45} \\ &= \left(\frac{N^*}{N}\right)^{a_N} \exp\left(b_N(N - N^*)\right), \tag{4.46} \end{align}$$

as N has a Gamma prior with parameters $a_N = b_N = 0.1$ (see Section 4.4.1).

Gamma shape parameter $\boldsymbol{\alpha}$

New values α_g^* are proposed from a lognormal random walk centered at the current value, i.e.

$$\alpha_g^* | \alpha_g \sim LN(\alpha_g, v_\alpha), \tag{4.47}$$

with $v_\alpha = 0.1$ which proved to provide good mixing. The acceptance probability will therefore be $\min(1, A)$ where

$$\begin{aligned} A &= \frac{\pi(\alpha_g^*)}{\pi(\alpha_g)} \times \frac{\pi(D|\theta^*, \tau)}{\pi(D|\theta, \tau)} \times \frac{q(\alpha_g|\alpha_g^*)}{q(\alpha_g^*|\alpha_g)} && (4.48) \\ &= \exp(\alpha_g - \alpha_g^*) \times \frac{\pi(D|\theta^*, \tau)}{\pi(D|\theta, \tau)} \times \frac{\alpha_g^*}{\alpha_g}, && (4.49) \end{aligned}$$

as the prior distribution for α_g is an Exponential with rate 1.

5

Analysis of Simulated Data

Computer simulations are useful as they can characterize the expected performance of phylogenetic methods under idealized conditions. Although it may not be possible to simulate data under models representative of reality, it certainly is possible to simulate data under the conditions assumed by the model. It is possible, then, to examine the performance of methods under best-case conditions (i.e., when all the assumptions of the model are met).

The general approach taken in this study was to construct DNA sequence data for a specific ordered species tree, under our model assumptions using computer simulation.

5.1 Data Simulation

A known rooted twelve-taxon tree was used as the model species tree S in this study. Four multiple sequence alignments, one for each gene, were generated under our model assumptions, in the following way:

- An order o was assigned to the vertices of S at random, giving S_o (Figure 5.1).

- Conditional on this order, κ_g xSPRs, for $g = 1, \ldots, 4$, were simulated from our model for each gene (with $\kappa_1 = 0$, $\kappa_2 = 0, \kappa_3 = 1, \kappa_4 = 2$), giving 4 unrooted topologies (τ_1, τ_2, τ_3 and τ_4). Topologies τ_1 and τ_2 are unrooted versions of S_o since they suffered no perturbation while τ_3 and τ_4 are the results of the LGT histories $Y_3 = \{Y_{31}\}$ and $Y_4 = \{Y_{41}, Y_{42}\}$, as indicated in Figure 5.1.

- Lengths assigned to each topology's branches were drawn randomly from an Exponential distribution, with $\ell_{gi} \sim Exp(10)$, for gene g and branch i.

- For each gene tree $T_g = (\tau_g, \boldsymbol{\ell_g})$, a multiple sequence alignment was generated according to the HKY85+$\mathit{\Gamma}$ substitution model (see Chapter 2.1 for further details) with the following parameter choice:
  - $\rho_g = 1.5$
  - $\alpha_g = 0.7$
  - $\boldsymbol{\pi}_g = (0.15, 0.35, 0.2, 0.3)$ for nucleotides A, G, C, T, respectively.

Our model assumes an ordered rooted species tree in order to constrain the LGTs that generate each gene tree in such a way that they occur between contemporary species, but the resulting gene tree is assumed to be an unrooted tree in terms of likelihood calculation (as explained in Subsection 2.2.1) . The simulated S_o as well as the LGT events that originated τ_3 and τ_4 are shown in Figure 5.1, and the unrooted topologies τ_3 and τ_4 can be seen in Figures 5.2 and 5.3.

5.2 MCMC

Using our implementation of the model in Java, three replicate MCMC runs were performed. The initial values for each chain were the true value for all the parameters except $\boldsymbol{\kappa}$, $\boldsymbol{Y}$, $\boldsymbol{\tau}$ and $\boldsymbol{\ell}$. This was done in order to decrease the number of iterations

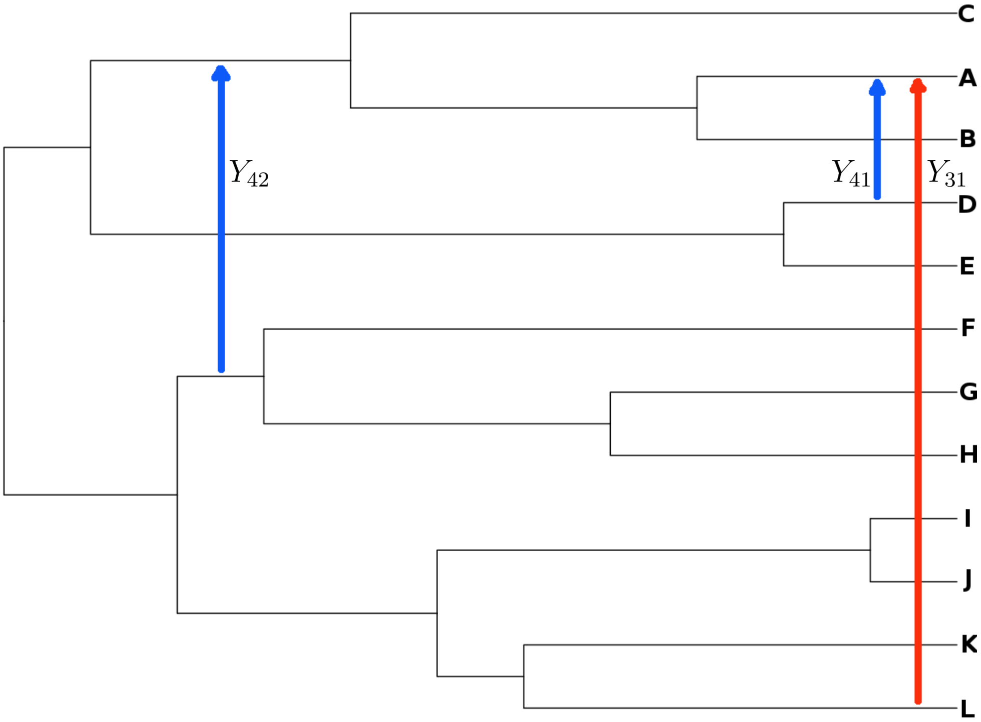


Figure 5.1: **Ordered species tree S_o and LGT histories for genes 3 and 4.** The tree topologies for genes 1 and 2 are identical to S_o, while the tree topology for gene 3 resulted from the LGT marked in red and that for gene 4 from the occurrence of LGTs represented in blue.

until the chain achieved convergence. Each initial topology was taken to be the topology of S_o, ($\tau_g^{(0)} = S_o$), with $\kappa_g = 0$ and $Y_g = \emptyset$, for $g = 1, \ldots, 4$. Each initial branch length ℓ_{gi} was drawn from the prior.

One million iterations were performed for each chain (at a rate of approximately 2000 iterations per hour in an Intel(R) Core(TM) i7 CPU 870 @ 2.93GHz with 4GB RAM) including a burn-in phase of 300K iterations. The chain was thinned every 100 iterations. Parameters were updated by using the methods described in Subsection 4.4.2. Gene tree topologies, LGT number and LGT history were updated by using both backward-forward and one-step backward-forward proposals. Branch lengths were updated using the single branch length proposal. The next section describes the results obtained.

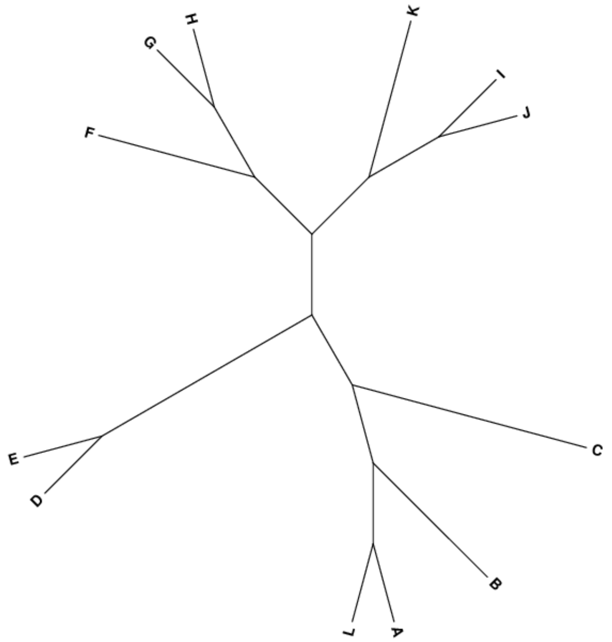


Figure 5.2: Unrooted gene tree topology for gene 3, τ_3.

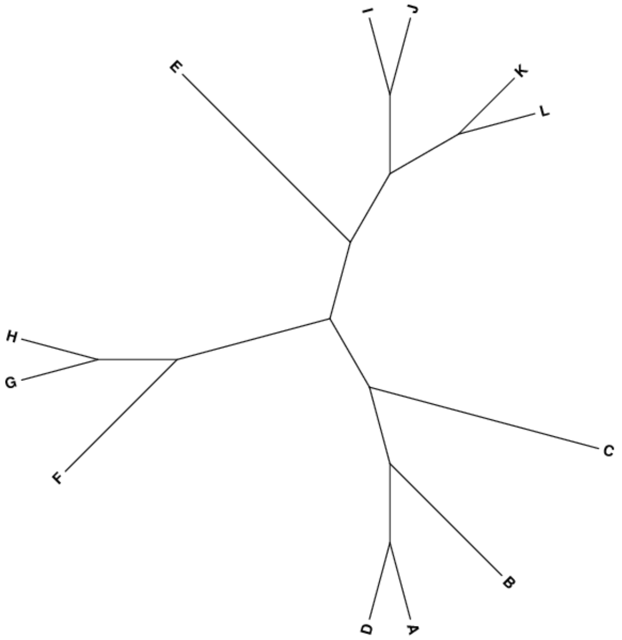


Figure 5.3: Unrooted gene tree topology for gene 4, τ_4.

5.3 Results

5.3.1 Overall look at MCMC convergence

As our model for this specific simulation includes more than 80 parameters (the number of LGT events in each LGT history $\boldsymbol{Y_g}$ is variable), in this subsection we will discuss only the MCMC trace, ACF and density plots for the three chains and some selected parameters which we believe are representative of the overall results.

Figure 5.4 displays the MCMC results for parameters α_g, ρ_g, π_{gA}, λ_g and ℓ_{g1} for gene 4. It can be observed that the three chains converged to the same distribution which gives strong support to the true value in all cases. No issues were found in terms of mixing or high autocorrelation. Similar results were achieved for genes 1, 2 and 3.

In Figure 5.5 we observe similar plots for the across-gene level parameters, all showing proper mixing and convergence again to the same posterior distribution, while Figure 5.6 shows the MCMC results for the number of LGT events, this time for all four genes. All chains converged to the same posterior distribution for genes 1, 2 and 3, but for gene 4, the green chain, after one million iterations converged to a different region in parameter space in comparison to the red and blue chains. This fact is of great importance and will be addressed in detail in Subsection 5.3.6.

As only red and blue chains reached the same posterior distribution for all parameters, in the next sections we will mainly discuss the posterior results obtained from the red chain, although the results of the three chains will be addressed whenever appropriate.

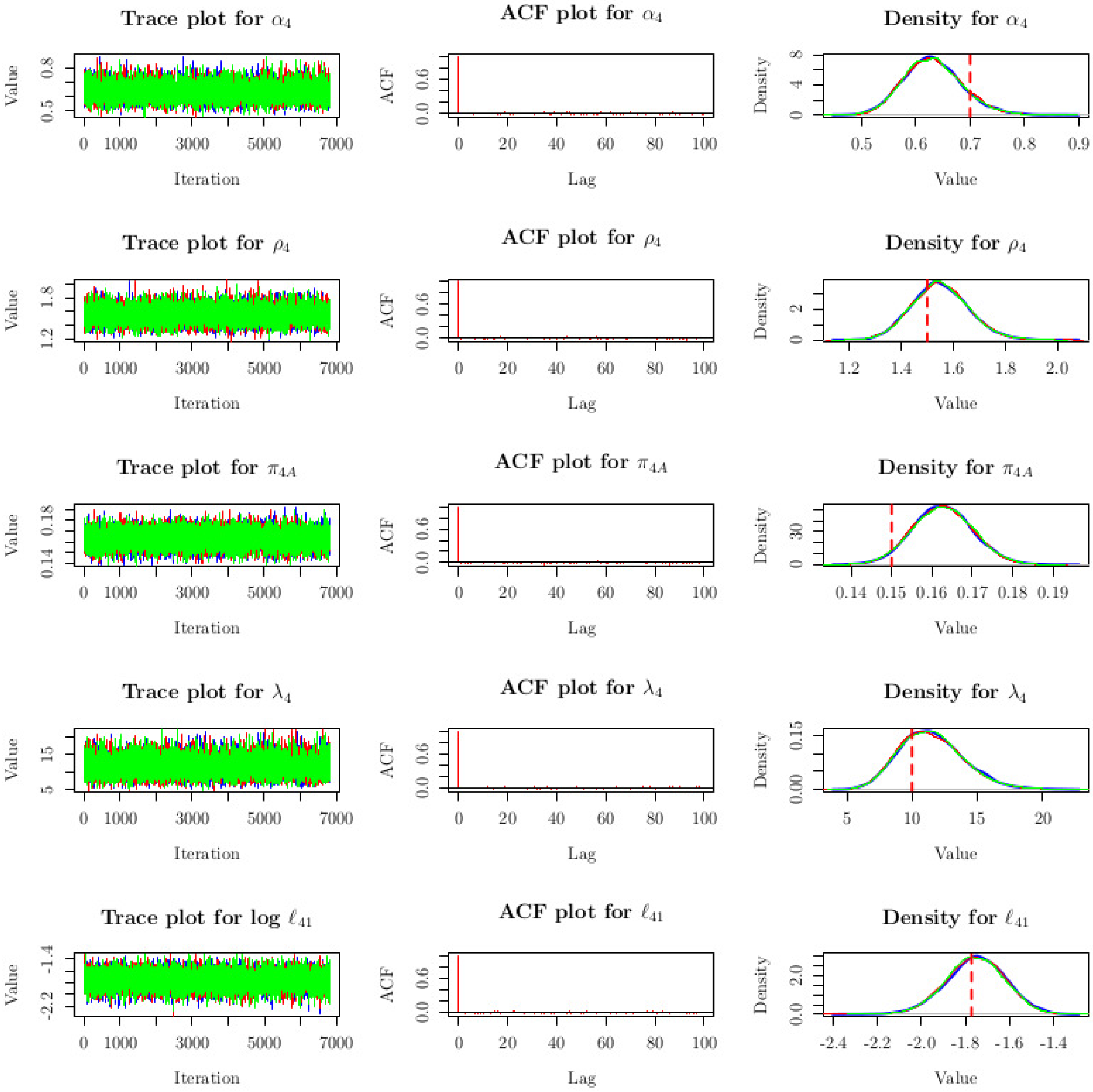


Figure 5.4: MCMC results for within-gene level parameters α_g, ρ_g, π_{gA}, λ_g and ℓ_{g1} for gene 4. The three chains are coloured green, red and blue and dashed lines denote the true parameter values.

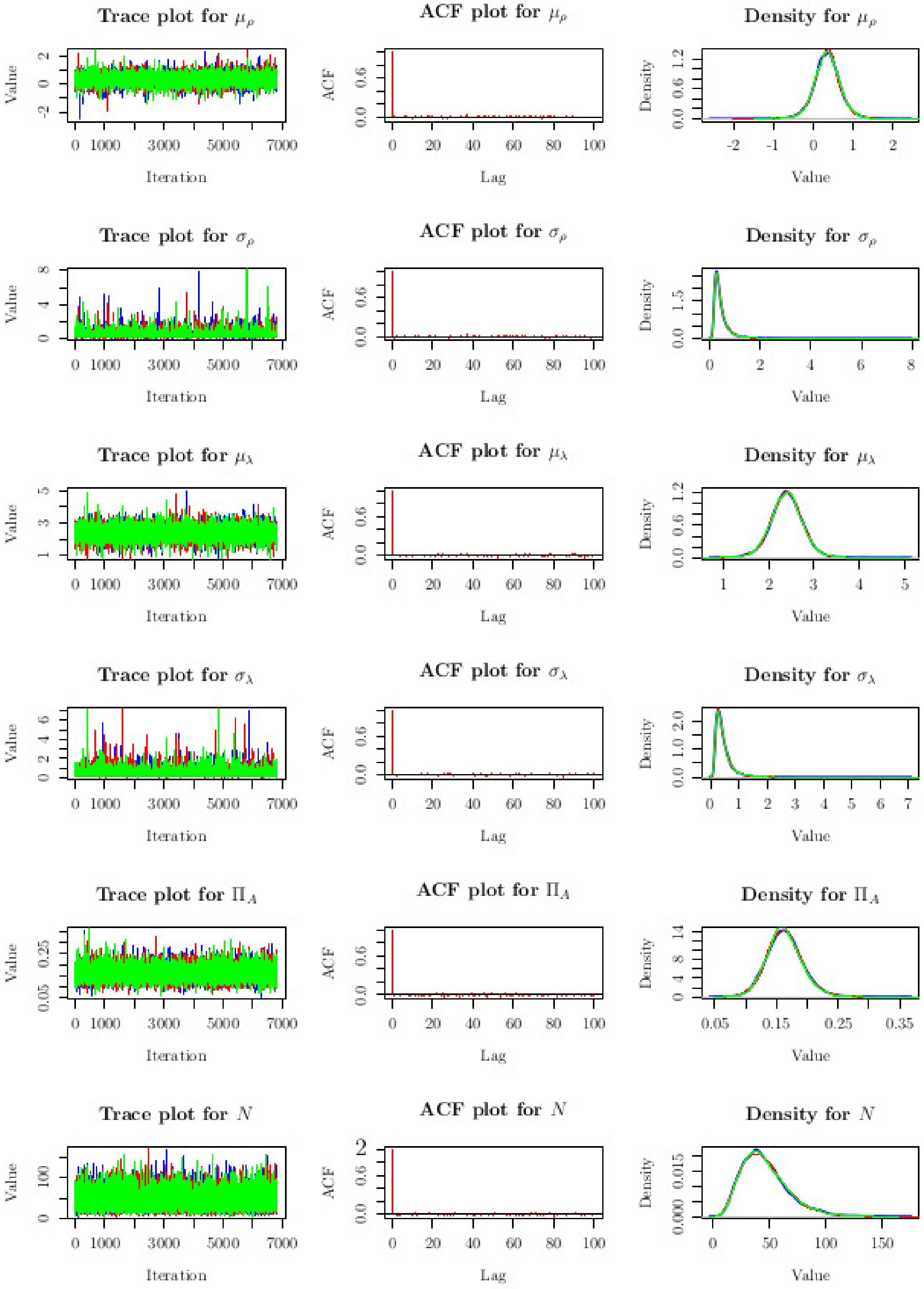


Figure 5.5: MCMC results for across-gene level parameters $\boldsymbol{\phi}$. The three chains are coloured green, red and blue.

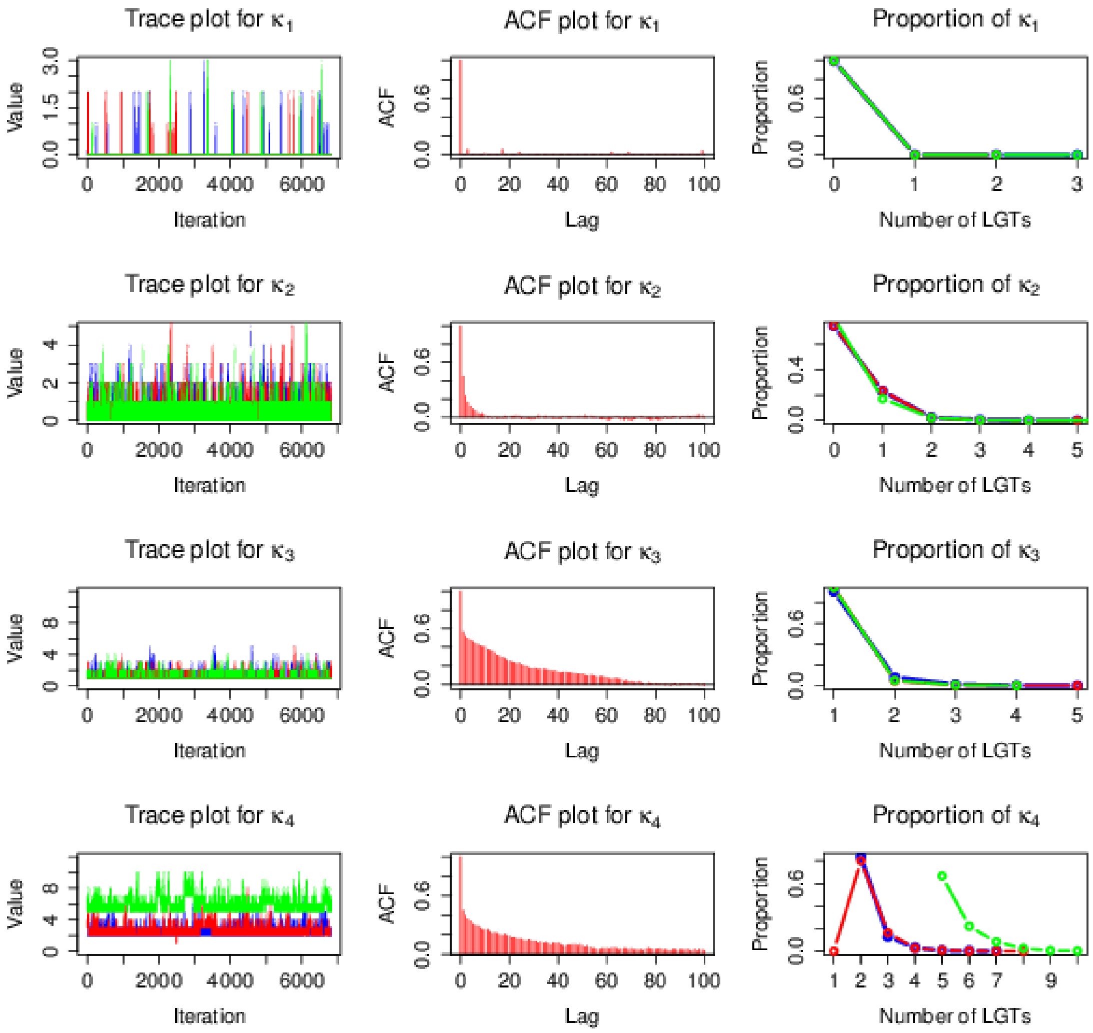


Figure 5.6: MCMC results for the number of LGT events $\boldsymbol{\kappa}$. The three chains are coloured green, red and blue.

5.3.2 Within-gene level parameters

Table 5.1 presents the results for the within-gene level parameters for the simulated genes. Listed in the table are the posterior mean and 95% HPD Bayesian credible intervals for each Gamma shape parameter α_g, transition-transversion ratio ρ_g, composition vector $\boldsymbol{\pi}_g$ and expected divergence λ_g; the table also contains the posterior means of the four $\boldsymbol{\pi}$ parameters. We see that the posterior means are close to the true values ($\alpha_g = 0.7$, $\rho_g = 1.5$, $\lambda_g = 10$), which are in all cases within the credible interval, except for ρ_2. We note that the true value of ρ_2 $(= 1.5)$ is in the tail of its posterior distribution, as we can see in Figure 5.7. This might be explained by the simulated sequence being analysed not being particularly consistent with the true value. Only with very long sequences would such an outlying point indicate something suspicious (in terms of the code or MCMC run).

The posterior means of the stationary distributions $\boldsymbol{\pi_g}$ are equally very similar to the true values.

				π_g			
g	α_g	ρ_g	λ_g	A	G	C	T
1	0.69(0.56,0.82)	1.57(1.35,1.80)	13.15(7.63,18.54)	0.15	0.36	0.19	0.30
2	0.68(0.58,0.78)	1.26(1.10,1.42)	9.38(5.87,13.47)	0.15	0.34	0.22	0.29
3	0.77(0.64,0.91)	1.43(1.25,1.64)	12.33(7.54,17.46)	0.16	0.34	0.19	0.32
4	0.63(0.53,0.74)	1.55(1.34,1.76)	11.52(6.90,16.22)	0.16	0.36	0.19	0.28

Table 5.1: Within-gene level parameters for the simulated data. For each gene g we have the posterior mean and 95% credible intervals for α_g, ρ_g and λ_g, and the posterior means for $\boldsymbol{\pi_g}$. The true values for each parameter are as follows: $\alpha_g = 0.7$, $\rho_g = 1.5$ and $\lambda_g = 10$, and $\boldsymbol{\pi_g} = (0.15, 0.35, 0.20, 0.30)$

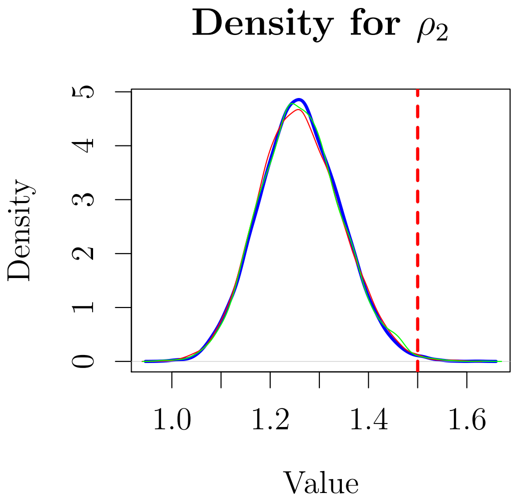


Figure 5.7: **Posterior distribution for $\boldsymbol{\rho_2}$ for all three chains.** Dashed red line represents the true value.

5.3.3 Across-gene level parameters

Table 5.2 lists the results for the across-gene level parameters used to pool information about $\boldsymbol{\rho}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$. The data were simulated using $\rho_g = 1.5$ and $\lambda_g = 10$, for $g = 1, \ldots, 4$, and so it is helpful to study the posterior distribution on this original scale for $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$ (rather than on the log scale of μ_ρ and μ_λ). Now

$$E(\rho_g|\mu_\rho, \sigma_\rho^2) \equiv \mu'_\rho = \exp\left(\mu_\rho + \frac{\sigma_\rho}{2}\right) \tag{5.1}$$

and

$$E(\lambda_g|\mu_\lambda, \sigma_\lambda^2) \equiv \mu'_\lambda = \exp\left(\mu_\lambda + \frac{\sigma_\lambda}{2}\right), \tag{5.2}$$

and so we will also look at the posteriors for μ'_ρ and μ'_λ.

The posterior means for these natural-scale parameters is 1.79 and 13.87 for μ'_ρ and μ'_λ respectively. We note that the values used for ρ_g and λ_g when simulating our data are included in the 95% credible intervals for μ'_ρ and μ'_λ. A similar reasoning can be applied to the central tendencies for the stationary distribution prior mean vector $\boldsymbol{\Pi}$. Table 5.2 also shows that the posterior means for $\boldsymbol{\Pi}$ are consistent with the composition vector values used to simulate the data ($\boldsymbol{\pi}_g = (0.15, 0.35, 0.20, 0.30)$ for $g = 1, \ldots, 4$).

Log-scale central tendencies		Natural-scale central tendencies		Measures of precision	
Param.	Mean (95% CI)	Param.	Mean (95% CI)	Param.	Mean (95% CI)
μ_ρ	0.36(-0.31,1.00)	μ'_ρ	1.79(0.91,3.39)	$1/\sigma^2_\rho$	2.27(0.98,11.11)
μ_λ	2.40(1.67,3.08)	μ'_λ	13.87(6.69,27.39)	$1/\sigma^2_\lambda$	2.17(0.93,9.09)
		$\mathit{\Pi}_A$	0.16(0.11,0.22)	$N_{\mathit{\Pi}}$	45.87(10.23,85.83)
		$\mathit{\Pi}_G$	0.34(0.26,0.41)		
		$\mathit{\Pi}_C$	0.20(0.14,0.26)		
		$\mathit{\Pi}_T$	0.30(0.22,0.36)		

Table 5.2: Across gene-level parameters for the simulated data. For each parameter we give the posterior mean and 95% credible intervals.

5.3.4 Gene Tree Topology ($\boldsymbol{\tau}$)

Table 5.3 summarizes the posterior probabilities for all topologies in the posterior distribution for each gene and each chain. The true topologies for each gene are the following:

1. ((C,(A,B)),(D,E),((F,(G,H)),((I,J),(K,L))));
2. ((C,(A,B)),(D,E),((F,(G,H)),((I,J),(K,L))));
3. ((C,(B,(A,L))),(D,E),((F,(G,H)),(K,(I,J))));
4. ((C,(B,(A,D))),(F,(G,H)),(E,((I,J),(K,L))));

For all genes, and all chains, the posterior mode corresponds to the true topology. In the case of the topologies for genes 1 and 2, as no LGTs were performed on S_o when generating either genes, the true topology is the species tree topology and their posterior probabilities are approximately 0.998 and 0.742, respectively. For genes 3 and 4, which were constructed by performing 1 and 2 LGTs on S_o respectively, the most probable topologies were the correct ones and had posterior probabilities around 0.983 and 0.987. Note that, despite the fact that the marginal posterior distribution of the green chain for κ_4 did not converge to a posterior distribution containing the true number of LGTs, its marginal distribution for τ_4 did reach a distribution containing only one topology, which is the true τ_4, thus this topology has a posterior probability of 1. This clearly indicates the multi-modal aspect of the LGT history parameter space which we will address later on.

5.3.5 Branch lengths ($\boldsymbol{\ell}$)

In relation to the branch lengths, Figure 5.8 shows the deviation of the posterior distribution from the true branch lengths for gene 4. Each boxplot summarises the difference log(posterior mean ℓ_{gi}) $-$ log(true ℓ_{gi}), for $i = 1, \ldots, B$ where B is the number of branches. The true branch length value is indicated on the x-axis and the branches have been ordered in increasing length. The boxplots clearly show that the variance of the deviation from the true value decreases as the true value increases, although the true value lies within the marginal 95% credible intervals for each branch length. Also, for the smallest branch length, the deviation from the true values is such that the true value appears in the tail of the posterior distribution. This reinforces the need for using a truncated prior which gives zero probability to branch lengths smaller than 0.002, as a branch length of 0.002 represents the occurrence of approximately 1 substitution every 500 nucleotides. As our simulated

	Gene 1			Gene 2			Gene 3			Gene 4		
Topology	Ch B	Ch R	Ch G	Ch B	Ch R	Ch G	Ch B	Ch R	Ch G	Ch B	Ch R	Ch G
((C,(A,B)),(D,E),((F,(G,H)),((I,J),(K,L))));	0.9984	0.9994	0.9999	0.7420	0.7429	0.8095	-	-	-	-	-	-
((D,E),(F,(G,H)),((C,(A,B)),((I,J),(K,L))));	0.0008	-	-	0.2104	0.2072	0.1549	-	-	-	-	-	-
((C,(A,B)),(F,(G,H)),((D,E),((I,J),(K,L))));	0.0007	0.0005	-	0.0475	0.0498	0.036	-	-	-	-	-	-
((C,(B,(A,L))),(D,E),((F,(G,H)),(K,(I,J))));	-	-	-	-	-	-	0.9825	0.9875	0.9869	-	-	-
((C,(B,(A,L))),(D,E),((I,J),(K,(F,(G,H)))));	-	-	-	-	-	-	0.0153	0.0104	0.0107	-	-	-
((C,(B,(A,L))),(D,E),(K,((F,(G,H)),(I,J))));	-	-	-	-	-	-	0.0009	0.0012	0.0011	-	-	-
(C,((B,(A,L)),(D,E)),((F,(G,H)),(K,(I,J))));	-	-	-	-	-	-	0.0007	0.0004	0.0005	-	-	-
((B,(A,L)),(C,(D,E)),((F,(G,H)),(K,(I,J))));	-	-	-	-	-	-	0.0006	-	0.0002	-	-	-
((C,(B,(A,L))),(D,E),((F,(G,H)),(J,(I,K))));	-	-	-	-	-	-	-	0.0003	0.0002	-	-	-
((C,(B,(A,L))),(D,E),((F,(G,H)),(I,(J,K))));	-	-	-	-	-	-	-	0.0002	0.0002	-	-	-
((C,(B,(A,D))),(F,(G,H)),(E,((I,J),(K,L))));	-	-	-	-	-	-	-	-	-	0.9868	0.9998	1
((C,(B,(A,D))),(F,(G,H)),((E,(I,J)),(K,L)));	-	-	-	-	-	-	-	-	-	0.0126	-	-
((C,(B,(A,D))),(E,(F,(G,H))),((I,J),(K,L)));	-	-	-	-	-	-	-	-	-	0.0006	-	-
(E,(C,(B,(A,D))),((F,(G,H)),((I,J),(K,L))));	-	-	-	-	-	-	-	-	-	-	0.0001	-

Table 5.3: Posterior distribution for gene tree topology by gene and MCMC run. The true topology for genes 1 and 2 is represented in red, while true topologies for genes 3 and 4 are represented in blue and green, respectively.

sequences are only 1000 nucleotides long, the data are unlikely to have enough information to infer very small branch lengths. Similar results were obtained for the genes 1, 2 and 3 (see Appendix B).

5.3.6 Number of LGTs ($\boldsymbol{\kappa}$)

Figure 5.9 displays the posterior distributions for the number of LGT events, κ_g, per gene and for each chain. All chains converged to the same distribution for genes 1, 2 and 3 and their posterior modes correspond to the true values ($\kappa_1 = 0$, $\kappa_2 = 0$ and $\kappa_3 = 1$). The posterior mode for genes 1 and 2 is zero with this value having posterior probabilities of approximately 0.997 and 0.742, respectively, in each chain. For gene 3, the posterior probability is concentrated on one LGT event with a probability of approximately 0.900 also, for all chains. For gene 4, and as we referred to before, while the blue and red chains converged to a posterior distribution with its mode at the true value (with a posterior probability of approximately 0.800 in both chains), the green chain has a posterior mode of 5, despite convergence to the correct gene tree topology (see Subsection 5.3.4). A possible explanation relates to the fact that the generation of a gene tree from an ordered species tree can be explained by several different LGT histories with different numbers of LGTs. Although our Geometric prior on κ_g favours LGT histories with smaller numbers of LGTs, the complexity of the parameter space often makes the transition between modes difficult, resulting in poor mixing. We believe this issue is related to the joint backward-forward proposal mechanism (see Chapter 4.4.2) for the LGT history, number of LGTs and gene trees.

The proposal we use is designed to use local moves to explore the tree space locally (see Figure 4.10). However, consider the following situation. Suppose during burn-in, the chain has entered a part of parameter space for which the correct gene tree topology has been found, but for which the corresponding sequence of LGTs is

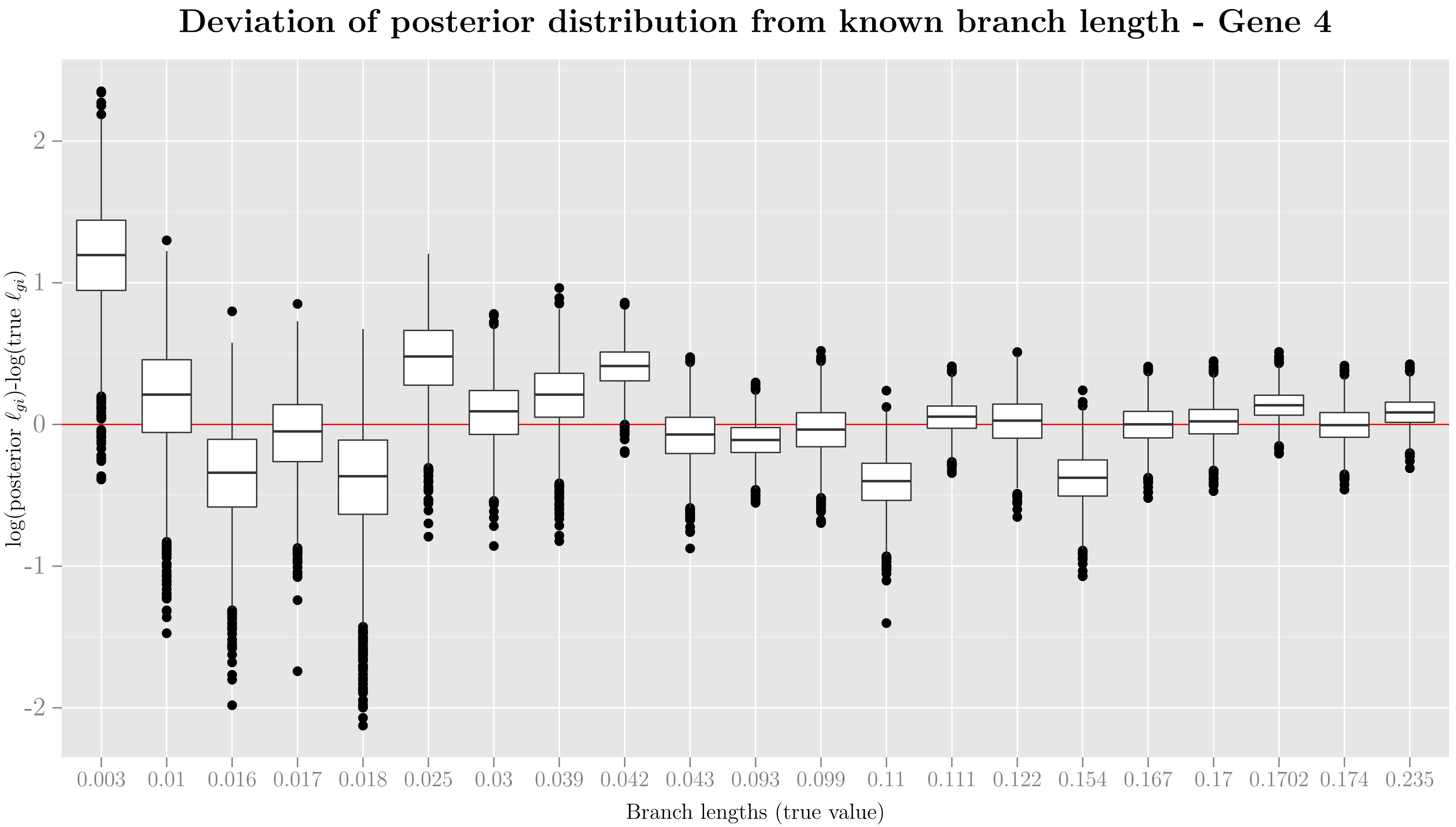


Figure 5.8: Difference between the log of the posterior values and the log of the true value for branch lengths for gene 4.

incorrect (in particular with κ_g too large). Then, backward-forward proposals which alter the gene tree topology will be heavily penalised due to a possibly large change in likelihood to an incorrect gene tree topology. Gross changes to the LGT history will only occur if the backward-forward proposal happens to propose an alternative LGT history that renders the correct gene tree topology. However, this is extremely unlikely to occur. Thus, the LGT history can get stuck in a long 'loop' between S_o and the correct gene tree topology, and the forward-backward proposal is very inefficient at 'tightening' such loops, since this can only be achieved within this (joint) proposal by changing the gene tree topology.

Our first attempt to address this issue was to use Metropolis-coupled MCMC (see Chapter 2.2.2 for more information) but no significant improvement occurred in the mixing of the chain. We could use an independence proposal, where, at each iteration, we would propose topologies from a random walk initialised always on S_o and with $\kappa_g \sim Po(a)$ steps, for known a, to allow global moves and thereby to explore other areas of the space but experimental results showed such a proposal to have an extremely low acceptance rate. Another solution, which brought better results was to develop another joint proposal, the one-step backward-forward proposal. A special case of the backward-forward proposal, it allows the chain to move only one step forward or one step backward, with only one step forward being allowed if the current state for topology is S_o. The idea here is that, by making even smaller local moves, the current state gene tree and LGT history are less disturbed and this should lead to increasing the acceptance probability.

Therefore, by allowing both backward-forward and one-step backward-forward samplers to work sequentially (a backward-forward move was proposed followed by a one-step backward-forward move proposal), we achieved better mixing in our MCMC analysis. Convergence issues relating to the LGT history occurred only for

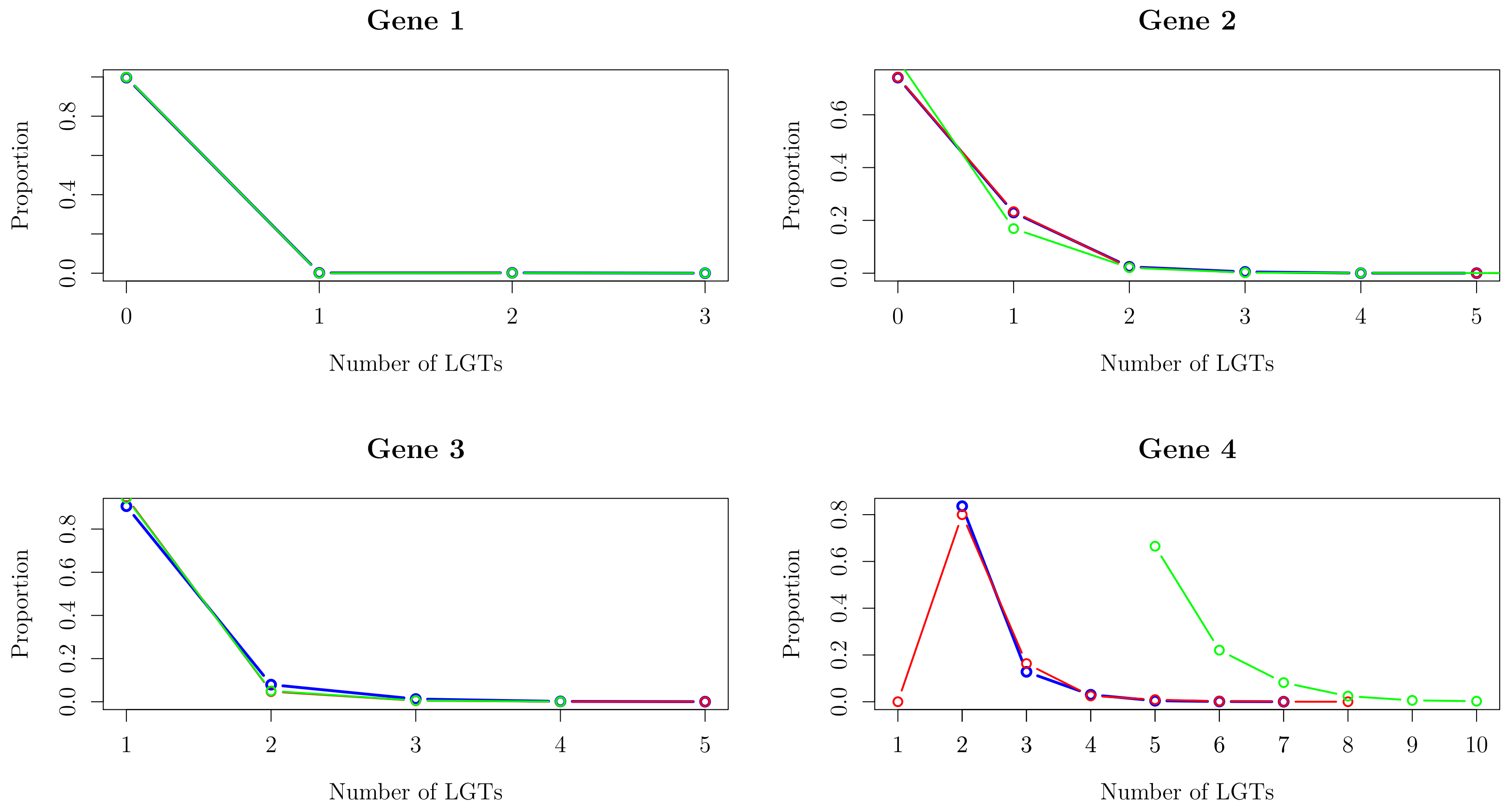


Figure 5.9: Number of LGTs for all genes and MCMC runs.

gene 4 (with $\kappa_4 = 2$), where around 70% of the chains converged to the true LGT history.

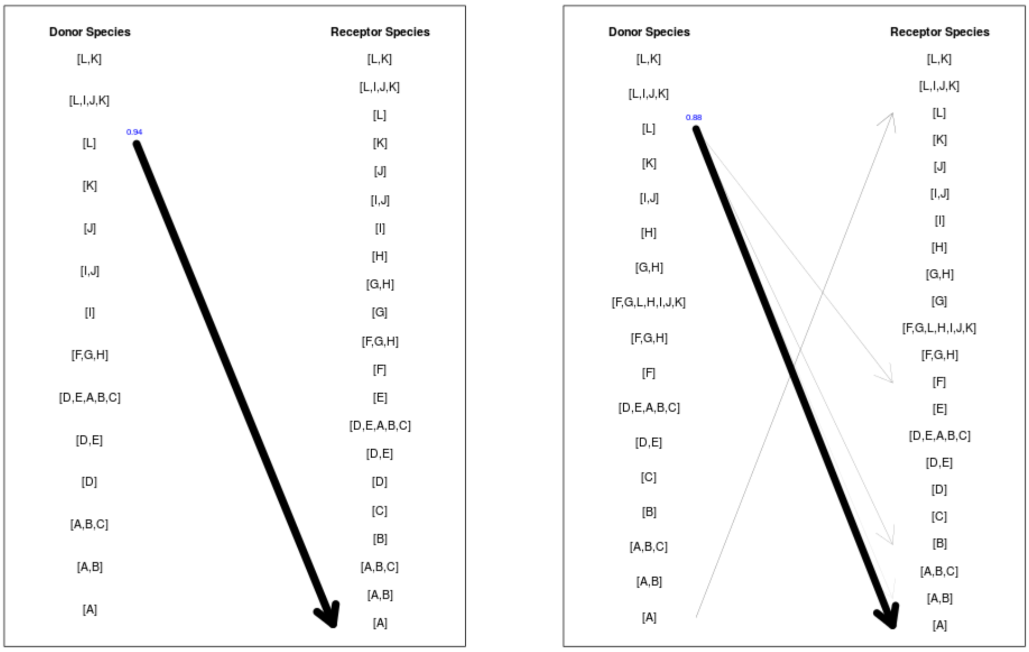
Our aim is to propose LGT histories between the current rooted species tree order state S_o and gene tree topology. This will form part of our future work (see Chapter 7).

5.3.7 LGT history ($\boldsymbol{Y}$)

To present the LGT history we will introduce a novel visualisation tool, the LGT Biplot, which represents the LGTs as arrows from the donor species to the receptor species. The thickness of the arrow is a function of the posterior probability - the thicker the arrow, the higher the probability. On the left is a list of "graft" splits, and on the right "prune" splits. The posterior probability is obtained by marginalising over all other aspects of the posterior distribution e.g. tree topology. The LGT Biplots in Figures 5.10 and 5.11 show the LGT events with non-negligible posterior probability (exceeding 0.05), for the red and blue chains, respectively. The true LGT history for genes 3 and 4 is represented in Figure 5.1 and can be also represented in terms of the edges or splits that participated in the event in the following way:

- **Gene 3:** [L] to [A];
- **Gene 4:** [D] to [A] and [F,G,H] to [A,B,C].

Note that, for gene 3, considering that this gene's evolutionary history resulted from only one LGT, and that we know the true τ_3, the only possible LGT that could originate it is [L] to [A]. Figure 5.10 shows us that our model has successfully recovered the true LGT history with posterior probabilities of 0.94 and 0.88 for the occurrence of the LGT from [L] to [A], for chains red and blue, respectively.



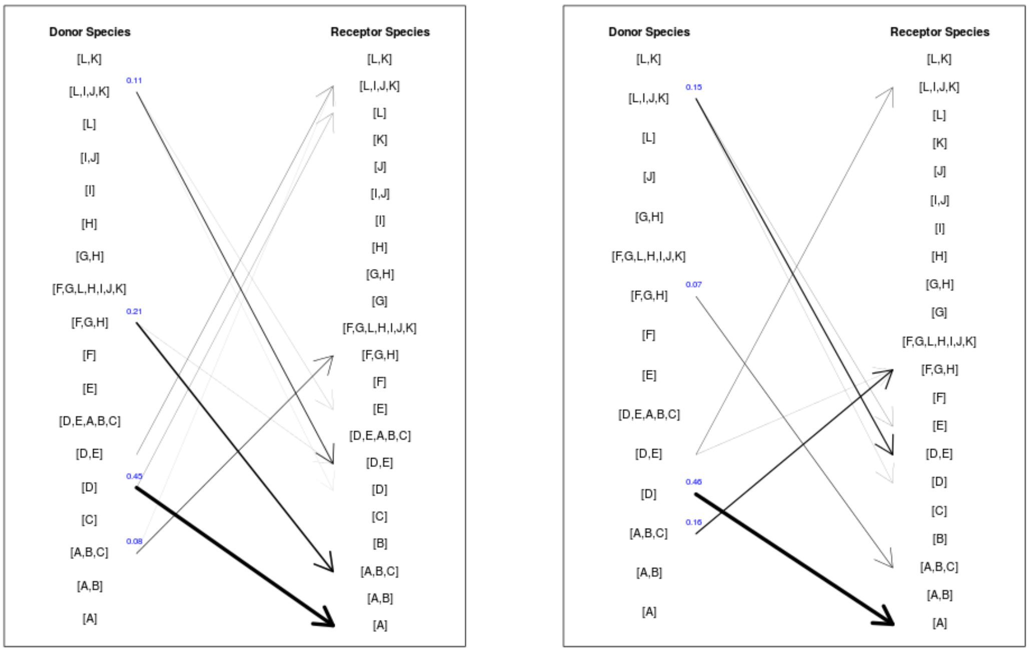
Gene 3 - Red Chain

Gene 3 - Blue Chain

Figure 5.10: Representation of the LGTs present in the posterior distribution for Gene 3.

A different situation characterises the LGT history for gene 4. As we already saw when analysing the inference results for κ_4, τ_4 can be derived from S_o through LGT histories with different numbers of LGTs. Nevertheless, due to the nature of the true LGTs (used to simulate the data), τ_4 can be generated through different LGT histories even with the same number of LGTs. When looking at Figure 5.11, we notice that although the LGT from [D] to [A] is undoubtedly the shortest path for attaching the branch connected to D to the branch connected to A, the other rearrangement on the gene tree topology can be originated via 4 different LGTs.

Our model assumes an ordered rooted species tree when modelling the LGTs that generates each gene tree, but the gene tree is assumed to be an unrooted tree in terms of likelihood calculation. Therefore, for our specific case, because the second LGT occurred between two edges connected to the edge containing the root, LGTs



Gene 4 - Red Chain

Gene 4 - Blue Chain

Figure 5.11: Representation of the LGTs present in the posterior distribution for Gene 4.

[F,G,H] to [A,B,C], [A,B,C] to [F,G,H] , [L,I,J,K] to [D,E] and [D,E] to [L,I,J,K] will give exactly the same gene tree topology. See Figure 5.12 for a visual explanation.

Therefore, while the posterior distributions of both chains agree in the occurrence of an LGT between [D] and [A], with posterior probabilities of 0.45 and 0.46, the same does not occur in relation to the second LGT. The LGT between [F,G,H] and [A,B,C] has posterior probabilities of 0.21 for the red chain and 0.07 for the blue chain. On the other hand, the LGT events [A,B,C] to [F,G,H] and [L,I,J,K] to [D,E] have probabilities of 0.16 and 0.15, respectively, in the blue chain, while their probabilities in the red chain were 0.08 and 0.11. The LGT between [D,E] to [L,I,J,K] is also present in both posterior distributions with probability below 0.05 in each case.

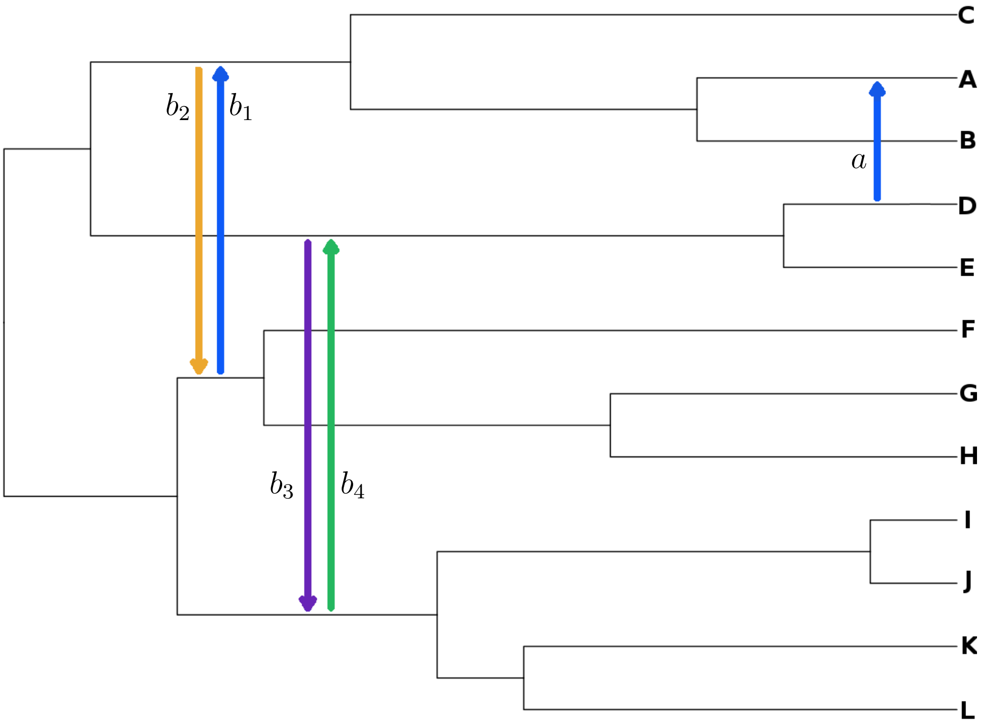


Figure 5.12: **Representation of 4 possible LGT histories with $\kappa_4 = 2$ LGTs for Gene 4.** Blue arrows a and b_1 indicate the LGTs that generated τ_4 in the simulated data, but any of the combinations of a with any b_i, for $i = 1, \ldots, 4$, gives the same unrooted gene tree topology represented in Figure 5.3.

6

Evolution of 2009 Swine-Origin Influenza Type A virus

6.1 Background

The influenza virus is an RNA virus of the family *Orthomyxoviridae*, which has been isolated from a wide range of hosts including humans, birds, pigs, horses and sea mammals. Its genome contains eight segments of single-stranded, negative-sense RNA. Three segments encode the polymerase complex: basic polymerase 2 (PB2), basic polymerase 1 (PB1) and the acidic protein (PA). The nucleoprotein segment (NP) encodes a protein that binds to viral RNA. The matrix segment (MP) encodes two proteins: a structural component of the viral capsid and a membrane ion channel. The non structural segment (NS) encodes a protein essential for cellular RNA processing and transport. Two other segments, hemagglutinin (HA) and neuraminidase (NA), encode viral surface glycoproteins responsible for host cell entry and exit, respectively. Based on the antigenicity of these two molecules, they are classified into 16 HA subtypes (H1-H16) and 9 NA subtypes (N1-N9) (Neumann

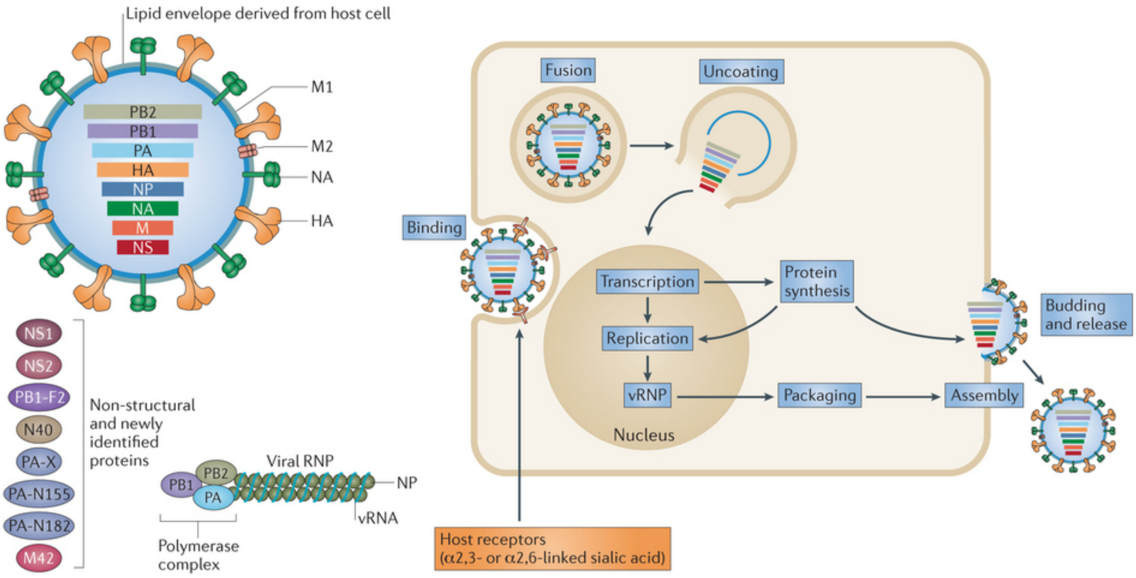


Figure 6.1: **Structure and life cycle of influenza A viruses**. Influenza A viruses are enveloped, single-stranded, negative-sense RNA viruses that contain eight gene segments that encode 16 proteins (Shi *et al.*, 2014).

et al., 2009).

When a cell is infected with an influenza virus the individual RNA segments enter the nucleus where they will be replicated. The new RNA segments are exported to the cytoplasm and incorporated into new virus particles which will be released from the cell (see Figure 6.1). If two or more influenza viruses infect the cell simultaneously, the RNAs of both viruses are replicated in the nucleus, promoting the assembling of new virus particles with 8 RNA segments originated from either infecting virus (Figure 6.2). This process is known as reassortment which is a form of lateral gene transfer. These reassortment events, associated with point mutations and inter-species transmission, can contribute to the emergence of new variants or strains with epidemic or pandemic potential. Pandemics are typically caused by the introduction of a virus with an HA subtype new to human populations.

In the twentieth century, three influenza viruses emerged in humans to cause major pandemics: the 1918 Spanish flu virus (H1N1), the 1957 Asian flu virus

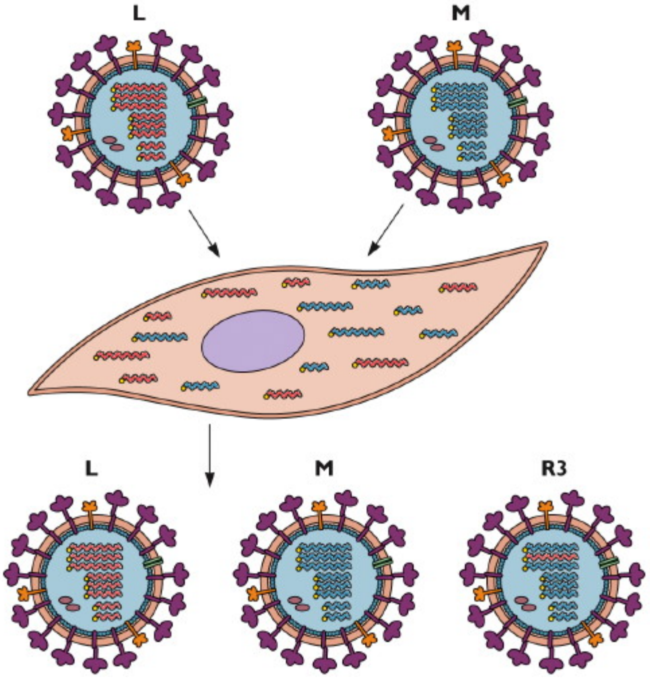


Figure 6.2: **Influenza virus reassortment.** This diagram shows a cell that is co-infected with two influenza viruses L and M. The infected cell produces both parental viruses as well as a reassortant R3 which inherits one RNA segment from strain L and the remainder from strain M (Racaniello, 2013).

(H2N2), and the 1968 Hong Kong flu virus (H3N2). These pandemics were initiated by the introduction and successful adaptation of a novel hemagglutinin subtype to humans from an animal source. These viruses later become established in humans as the cause of seasonal flu for many years until being replaced by a new pandemic virus.

A new swine-origin influenza A (H1N1) virus (S-OIV) emerged in Mexico and the United States in March 2009, spreading worldwide by human-to-human transmission and originating the first influenza pandemic of the twenty-first century (Smith *et al.*, 2009; Trifonov *et al.*, 2009; Kingsford *et al.*, 2009). As with most seasonal influenza viruses, this new virus is associated with a mild illness in the majority of people, although it is responsible for severe complications in more susceptible individuals.

On the basis of sequence similarity to previously reported swine influenza isolates, initial genetic characterization of the 2009 S-OIV outbreak suggested it had its origin in pigs in which the virus had been circulating for at least 80 years. A new triple-reassortant H3N2 virus, comprising genes from classical swine H1N1, North American avian, and human H3N2 influenza, was reported in 1998 as the cause of outbreaks in North American swine. In Europe, an avian H1N1 virus was introduced to pigs (avian-like swine H1N1) and first detected in Belgium in 1979. It is noteworthy that, until 2009, there was no evidence of Eurasian avian-like swine H1N1 circulating in North American pigs.

Using phylogenetic analyses, Smith *et al.* (2009) estimated a temporal reconstruction of the complex reassortment history of the 2009 S-OIV outbreak, summarized in Figure 6.3. They compared two genomes from the S-OIV outbreak with 811 genomes representing the spectrum of influenza A diversity (285 humans, 100 swine and 411 avian isolates) and constructed phylogenetic trees for each genomic sequence independently. Phylogenetic trees were inferred using the neighbour-joining distance method, with genetic distances calculated by maximum likelihood under the HKY85+Γ model. The parameters of this model were estimated using maximum likelihood on an initial tree. Further analyses were made to infer temporal phylogenies and rates of evolution.

Their results suggest that the S-OIV likely resulted from the reassortment of recent North American H3N2 and H1N2 swine viruses (i.e., avian/human/swine triple reassortant viruses) with Eurasian avian-like swine viruses. It also showed that each segment of the S-OIV genome seems to be nested within a well-established swine influenza lineage (which circulated in swine for $>$ 10 years before the 2009 outbreak), emphasising that the progenitor of the S-OIV epidemic had its origin in pigs.

Despite the fact that the precise evolutionary pathway of S-OIV origin is greatly hindered by the lack of surveillance data, these results seem to indicate that the polymerase genes, plus HA, NP and NS, emerged from a triple-reassortant virus circulating in North American swine. The source triple-reassortant itself comprised genes derived from avian (PB2 and PA), human H3N2 (PB1) and classical swine (HA, NP and NS) lineages. In contrast, the NA and M gene segments have their origin in the Eurasian avian-like swine H1N1 lineage (Smith *et al.*, 2009).

6.2 Analysis

We now try to infer the 2009 S-OIV evolution/reassortment events by using our model and the same multiple sequence alignments as Smith *et al.* (2009) (available at http://tree.bio.ed.ac.uk/people/arambaut/). As our model assumes a known species tree, we fixed the species tree to be the inferred tree for gene NP which, according to Smith *et al.* (2009), has been only vertically transmitted throughout the swine lineage. Due to the complexity of our model, and the computational and

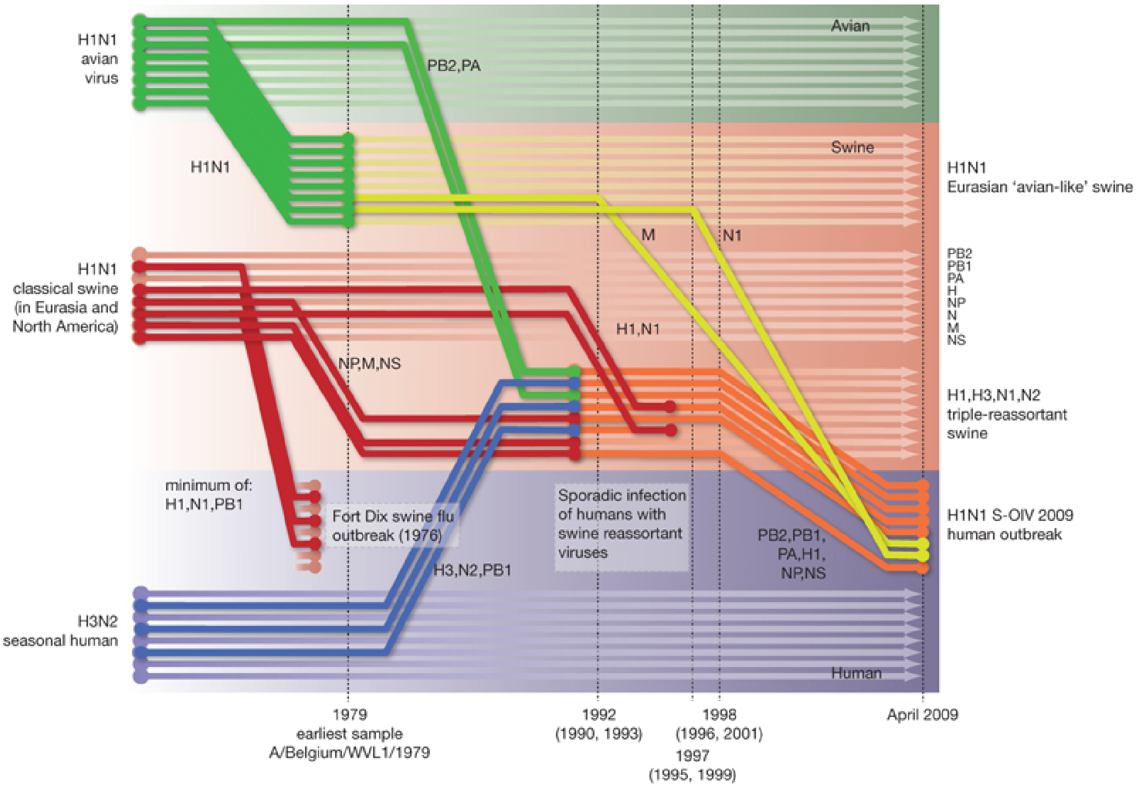


Figure 6.3: **Host species are represented by the different coloured shaded boxes: avian (green), swine (red) and human (blue).** The eight genomic segments are represented as parallel lines in descending order of size. Dates marked with dashed vertical lines indicate the mean time of divergence of the S-OIV genes from corresponding virus lineages. Reassortment events not involved with the emergence of human disease are omitted (Smith *et al.*, 2009).

time constraints, we sampled 26 virus representatives of eight virus lineages that were likely involved in the reassortment events that gave origin to the 2009 S-OIV (see Table 6.1). Species were chosen such that they would cover all lineages as well as different clades within the lineages. Most lineages are represented by 2 or 3 species, except the avian (with 8 species), reflecting the fact that Smith *et al.* (2009) used a higher number of avian species when compared to other lineages.

Our model assumes also that all genes are present in all taxa in the analysis and so we decided to remove genes HA and NA from our analysis as they are used to classify the different subtypes. As an example, if we choose to analyse viruses with gene H1 we would have to exclude some important lineages, e.g. the triple-reassortant swine H3N2 and H3N1. The species tree is represented in Figure 6.4.

Reference in Smith *et al.* (2009)	**Code**
4273_H3N2_Human_florida_ur070101_2008	A
2823_H3N2_Human_queensland_39_2003	B
3477_H3N2_Human_memphis_1_71	C
2693_H2N2_Human_albany_8_1967	D
2637_H2N2_Human_czechrepublic_1_1966	E
2094_H1N1_Human_denver_57	F
2471_H1N1_Human_oregon_ur060291_2007	G
2601_H1N1_Human_wilsonsmith_33	H
36_H1N1_Swine_swine_ohio_23_1935	I
46_H1N1_Swine_swine_wisconsin_2_1970	J
112_H1N1_Swine_swine_iowa_1_1986	K
155_H3N2_Swine_swine_manitoba_12707_2005	L
150_H3N1_Swine_swine_in_pu542_04	M
141_H1N2_Swine_swine_shanghai_1_2007	N
00_Canada_ON_RV1527_2009 (H1N1)	O
2593_H1N1_Human_california_04_2009	P
32_H1N1_Swine_swine_chonburi_niah9469_2004	Q
135_H1N1_Swine_swine_england_wvl16_1998	R
811_H5N1_Avian_chicken_shanxi_2_2006	S
356_H1N1_Avian_duck_italy_69238_2007	T
1713_H9N2_Avian_chicken_gansu_2_99	U
731_H4N8_Avian_duck_victoria_5384_2002	V
576_H3N8_Avian_redneckedstint_westernaustralia_4923_1983	W
372_H1N1_Avian_quail_in_38685_1993	X
562_H3N6_Avian_mallard_maryland_1235_2006	Y
358_H1N1_Avian_pigeon_mn_1407_1981	Z

Table 6.1: Taxa included in our phylogenetic analysis and corresponding reference in Smith *et al.* (2009). Each sample was coded to an alphabetical letter to facilitate the presentation of results.

Similarly to the analysis of simulated data in Chapter 5, we initialised three MCMC runs (which we will identify throughout this chapter as blue, red and green) with the following parameter choice:

- $\rho_g = 1.5$
- $\alpha_g = 0.7$
- $\boldsymbol{\pi}_g = (0.15, 0.35, 0.2, 0.3)$ for nucleotides A, G, C, T,

respectively. Also, the initial vertex order for S_o was assigned uniformly at random (Figure 6.4) and the initial gene tree topologies were taken to be the topology of S_o. The initial branch lengths were drawn at random from an Exp(10) distribution. Next, we present the MCMC results of iterations 3,000-265,000, which took around 11 weeks of computational time to obtain.

6.3 Results

In the next subsections we discuss the results for the most relevant parameters in this analysis.

6.3.1 Within and between gene-level parameters

All within and between gene-level parameters converged to the same marginal posterior distributions in each MCMC chain. Table 6.2 lists the posterior mean and Bayesian confidence intervals for the Gamma shape parameters α_g, transition-transversion ratios ρ_g and expected divergences λ_g, as well as the posterior means for $\boldsymbol{\pi}_g$.

Although in general, for all the genes, the nucleotide sites seem to evolve at a very low rate, the most noticeable feature of the table is that the posterior mean of α_g, for gene NS ($\alpha_{NS} = 0.37$), is almost twice the value for the other genes. This difference

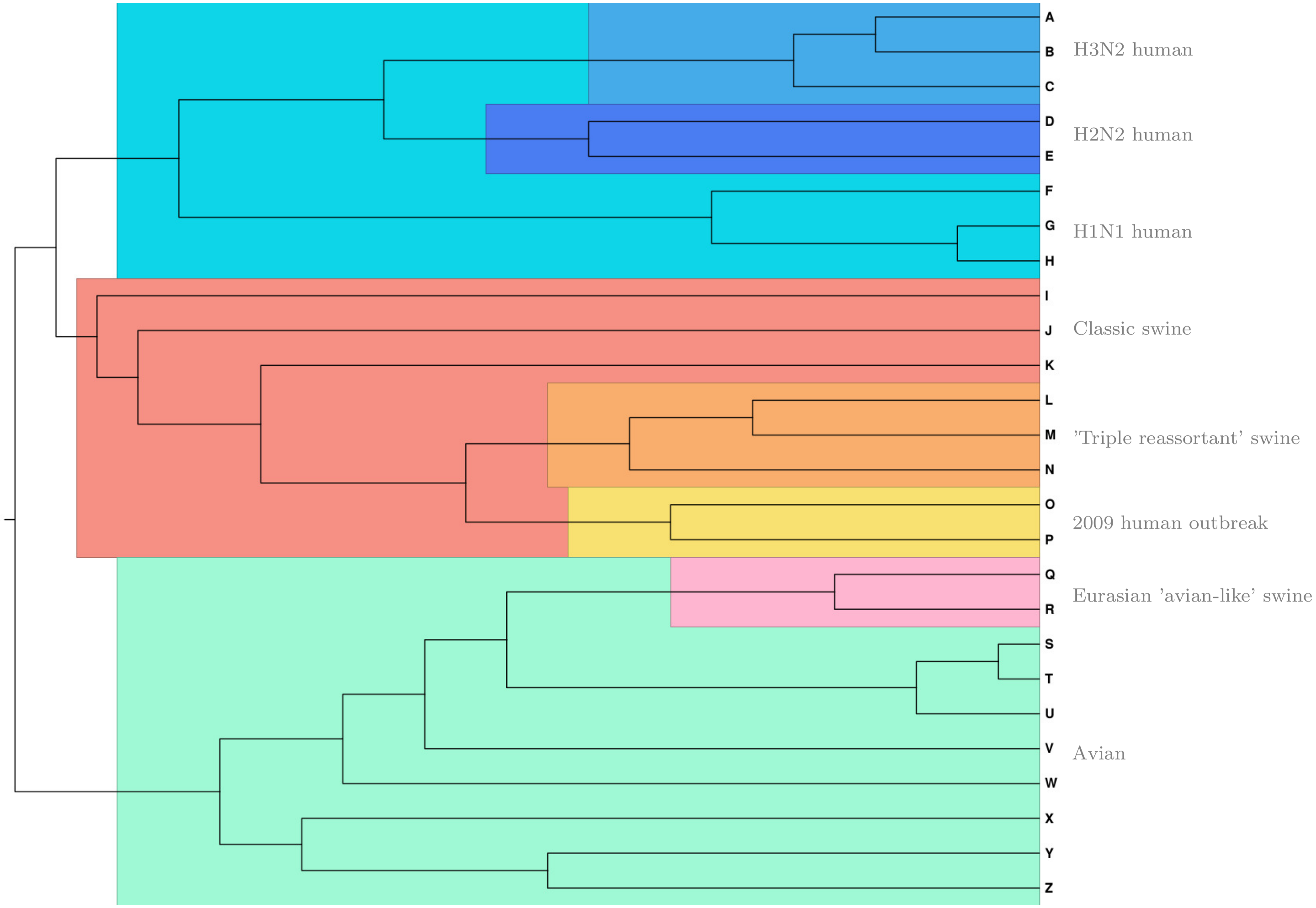


Figure 6.4: **Species tree relating 26 species of influenza Type A virus.** Clade colours and labels indicate major virus lineages and vertex order corresponds to the initial order assigned to S_o.

				π			
Gene	α	ρ	λ	A	G	C	T
PB2	0.20(0.18,0.22)	14.67(13.22,16.17)	22.54(16.35,28.68)	0.36	0.21	0.20	0.24
PB1	0.19(0.17,0.20)	14.07(12.51,15.41)	21.95(15.72,28.07)	0.35	0.20	0.21	0.24
PA	0.20(0.18,0.22)	14.15(12.55,15.65)	23.45(16.76,30.16)	0.35	0.21	0.20	0.23
NP	0.21(0.19,0.24)	10.73(9.46,12.02)	25.47(18.27,32.98)	0.34	0.22	0.21	0.23
M	0.19(0.16,0.23)	10.64(8.72,12.63)	**34.04(24.14,44.38)**	0.32	0.24	0.21	0.23
NS	**0.37(0.31,0.44)**	**7.17(6.12,8.37)**	23.35(16.54,30.46)	0.34	0.22	0.20	0.24

Table 6.2: Within-gene level parameters for the six genes in this study. We present the posterior mean and 95% credible intervals for α, ρ and λ, and the posterior means for $\boldsymbol{\pi}$.

is supported by the 95% CI (0.31,0.44) suggesting that it has a larger number of sites evolving at a higher rate than the other genes. On the other hand, the posterior mean of the transition/transversion rate ratio for the same gene ($\rho_{NS} = 7.17$) is distinguishably smaller than for the other genes, being half the value for genes PB2, PB1 and PA. The 95% CI (6.12,8.37) also supports this conclusion. The results for the across-gene level parameters are presented in Table C.1.

6.3.2 Topologies

The unrooted gene tree topologies with posterior probability higher than 0.1 are listed in Table 6.3 for all six genes and MCMC runs. For PB2, all 3 chains converged to a posterior distribution with a common topology as the posterior mode, with posterior probabilities for the modal topology ranging from 0.75 to 0.96. For gene NP, the red and green chains also converged to posteriors sharing a common supported topology although their posterior probabilities vary (0.95 and 0.43, respectively). The chains for the other four genes, at the time of the MCMC data collection, did not converge to the same posterior distribution, although the topologies are in most cases similar. The following results summarize the MCMC output even though clearly these chains have not fully converged.

	PB2			PB1			PA		
Topology	B	R	G	B	R	G	B	R	G
(H,(G,(F,(E,(D,(C,(B,A)))))),((I,(K,J)),(((T,W),((R,Q),(V,(U,S)))),(X,((Z,Y),((O,P),(N,(M,L))))))));	0.96	0.93	0.75	-	-	-	-	-	-
(H,(G,(F,(E,(D,(C,(B,A)))))),((I,(K,J)),(((T,W),((R,Q),(V,(U,S)))),((X,(Z,Y)),((O,P),(N,(M,L)))))));	-	-	0.17	-	-	-	-	-	-
((E,D),((Z,(X,Y)),((K,J),(I,(H,(F,G))))),((V,(W,((R,Q),(U,(T,S))))),(C,((O,P),((M,L),(N,(B,A)))))));	-	-	-	0.70	-	-	-	-	-
((E,D),((Z,(X,Y)),((K,J),(I,(H,(F,G))))),((V,(W,((R,Q),(U,(T,S))))),(C,((B,A),((O,P),(N,(M,L)))))));	-	-	-	0.16	-	-	-	-	-
((Z,(X,Y)),((C,(E,D)),((O,P),((M,L),(N,(B,A))))),(((K,J),(I,(H,(F,G)))),(W,(V,((R,Q),(U,(T,S)))))));	-	-	-	-	0.98	-	-	-	-
((W,((R,Q),(V,(U,(T,S))))),((H,(F,G)),(I,(K,J))),(((E,D),(Z,(X,Y))),(C,((B,A),((O,P),(N,(M,L)))))));	-	-	-	-	-	0.59	-	-	-
((C,((O,P),((M,L),(N,(B,A))))),((E,D),(Z,(X,Y))),(((H,(F,G)),(I,(K,J))),((W,V),((R,Q),(U,(T,S))))));	-	-	-	-	-	0.39	-	-	-
((I,(K,J)),(H,((F,G),(E,(D,(C,(B,A)))))),(Z,((Y,(R,Q)),((X,((O,P),(N,(M,L)))),((U,V),(W,(T,S)))))));	-	-	-	-	-	-	0.88	-	-
((H,(G,(F,(E,(D,(C,(B,A))))))),(I,(K,J)),(Z,(Y,((R,Q),(((U,X),((O,P),(N,(M,L)))),((T,S),(W,V)))))));	-	-	-	-	-	-	-	0.38	-
((H,(G,(F,(E,(D,(C,(B,A))))))),(I,(K,J)),(Z,((Y,(R,Q)),(((U,X),((O,P),(N,(M,L)))),((T,S),(W,V))))));	-	-	-	-	-	-	-	0.38	-
((I,(K,J)),(H,(G,(F,((E,D),(C,(B,A)))))),(Z,((Y,(R,Q)),(((U,X),((O,P),(N,(M,L)))),((T,S),(W,V))))));	-	-	-	-	-	-	-	0.13	-
((I,(K,J)),((H,Z),(G,(F,(E,(D,(C,(B,A))))))),(Y,((R,Q),(((U,X),((O,P),(N,(M,L)))),((T,S),(W,V))))));	-	-	-	-	-	-	-	-	0.47
((H,Z),(G,(F,(E,(D,(C,(B,A)))))),((I,(K,J)),((Y,(R,Q)),(((U,X),((O,P),(N,(M,L)))),((T,S),(W,V))))));	-	-	-	-	-	-	-	-	0.41

	NP			M			NS		
Topology	B	R	G	B	R	G	B	R	G
(H,(G,(F,(E,(C,(D,(B,A)))))),((I,(J,(K,((O,P),(N,(M,L)))))),((Y,(Z,X)),(W,(V,((R,Q),(U,(T,S))))))));	0.56	-	-	-	-	-	-	-	-
(H,(G,(F,(E,(D,(C,(B,A)))))),((I,(J,(K,((O,P),(N,(M,L)))))),((Y,(Z,X)),(W,(V,((R,Q),(U,(T,S))))))));	0.40	-	-	-	-	-	-	-	-
(H,(G,(F,(E,(D,(C,(B,A)))))),((I,(J,(K,((O,P),(N,(M,L)))))),((Z,(X,Y)),(W,(V,((R,Q),(U,(T,S))))))));	-	0.95	0.43		-	-	-	-	-
(H,(G,(F,(E,(C,(D,(B,A)))))),((I,(J,(K,((O,P),(N,(M,L)))))),((Z,(X,Y)),(W,(V,((R,Q),(U,(T,S))))))));	-	-	0.51		-	-	-	-	-
((I,(J,(K,(N,(M,L))))),(H,(G,(F,((E,D),(C,(B,A)))))),((Y,(Z,X)),((W,(R,(Q,(O,P)))),(V,(U,(T,S))))));	-	-	-	0.84	-	-	-	-	-
((I,(J,(K,(N,(M,L))))),(H,(G,(F,((E,D),(C,(B,A)))))),((Y,(Z,X)),(W,((R,(Q,(O,P))),((U,V),(T,S))))));	-	-	-	-	0.43	-	-	-	-
((I,(J,(K,(N,(M,L))))),(H,(G,(F,((E,D),(C,(B,A)))))),((Y,(Z,X)),(W,(V,((R,(Q,(O,P))),(U,(T,S)))))));	-	-	-	-	0.12	-	-	-	-
((I,(J,(K,(N,(M,L))))),(H,(G,(F,((E,D),(C,(B,A)))))),((Y,(Z,X)),(W,((R,(Q,(O,P))),(U,(V,(T,S)))))));	-	-	-	-	0.10	-	-	-	-
((I,(J,(K,(N,(M,L))))),(H,(G,((D,(F,E)),(C,(B,A))))),((Y,(Z,X)),(W,((R,(Q,(O,P))),(U,(V,(T,S)))))));	-	-	-	-	-	0.51	-	-	-
((I,(J,(K,(N,(M,L))))),(H,((F,G),((E,D),(C,(B,A))))),((Y,(Z,X)),(W,((R,(Q,(O,P))),(U,(V,(T,S)))))));	-	-	-	-	-	0.41	-	-	-
((I,(J,((K,Q),((O,P),(N,(M,L)))))),(H,(G,(F,((E,D),(C,(B,A)))))),((Z,(X,Y)),(W,(V,((R,T),(U,S))))));	-	-	-	-	-	-	0.33	-	-
((I,(J,((K,Q),((O,P),(N,(M,L)))))),(H,(G,(F,((E,D),(C,(B,A)))))),((Z,(X,Y)),((R,(T,(U,S))),(W,V))));	-	-	-	-	-	-	0.18	-	-
((I,(J,((K,Q),((O,P),(N,(M,L)))))),(H,(G,(F,((E,D),(C,(B,A)))))),((X,(Z,Y)),(W,((R,V),(T,(U,S))))));	-	-	-	-	-	-	0.12	-	-
((I,(J,((K,Q),((O,P),(N,(M,L)))))),(H,(G,(F,((E,D),(C,(B,A)))))),((X,(Z,Y)),(W,(V,(R,(U,(T,S)))))));	-	-	-	-	-	-	-	0.85	-
((I,(J,((K,Q),((O,P),(N,(M,L)))))),(H,(F,(G,((E,D),(C,(B,A)))))),((X,(Z,Y)),(W,(V,(R,(U,(T,S)))))));	-	-	-	-	-	-	-	-	0.75

Table 6.3: Posterior probabilities for gene tree topology by gene and MCMC run. Blue, red and green chains are represented by letters B, R and G, respectively. All the topologies present in the posterior distributions for the six genes were distinct from S_o and unique within and between genes.

6.3.3 Number of LGTs

Figure 6.5 shows the posterior distribution for the number of LGTs for each gene in the study. As mentioned before, all chains converged to the same posterior distribution for the topology of gene PB2, but in terms of the number of LGTs, the green chain clearly converged to a different posterior with mode equal to 12 LGTs, while the other two chains converged to very similar posteriors with mode equal to 8 LGTs. This behaviour was also observed and analysed in our simulation results (see Chapter 5.3.6), and is a result of the multimodal aspect of the LGT history parameter space and the difficulty in moving between the different modes. For genes PB1, PA, M and NS, the lack of convergence to the same posterior is obvious, while for gene NP, all chains converged to very similar posterior distributions, with mode equal to 5 LGTs.

6.3.4 LGT history

In the sequence of the previous results, we will start by presenting and discussing the LGT history for gene PB2.

The LGT Biplots in Figure 6.6 show the posterior LGT events with corresponding posterior probability (only for $p \geq 0.05$), for chains red and blue, respectively. Both LGT histories are similar and in both, seven of the LGTs are equally probable with a posterior probability of 0.12, although the two sets of LGTs are only partially overlapping. The LGT between clade [F,G,H] and [D,E,A,B,C] is redundant causing no change in the topology and in both LGT histories only two LGTs are related with the 2009 S-OIV (blue chain: [Y,Z] to [L,M,N,O,P] and [A,B,C,D,E] to [L,M,N,O,P]; red chain: [Y,Z] to [L,M,N,O,P] and [F,G,H] to [L,M,N,O,P]). These LGTs are represented on the species tree in Figure 6.7. All the other LGT events, which are

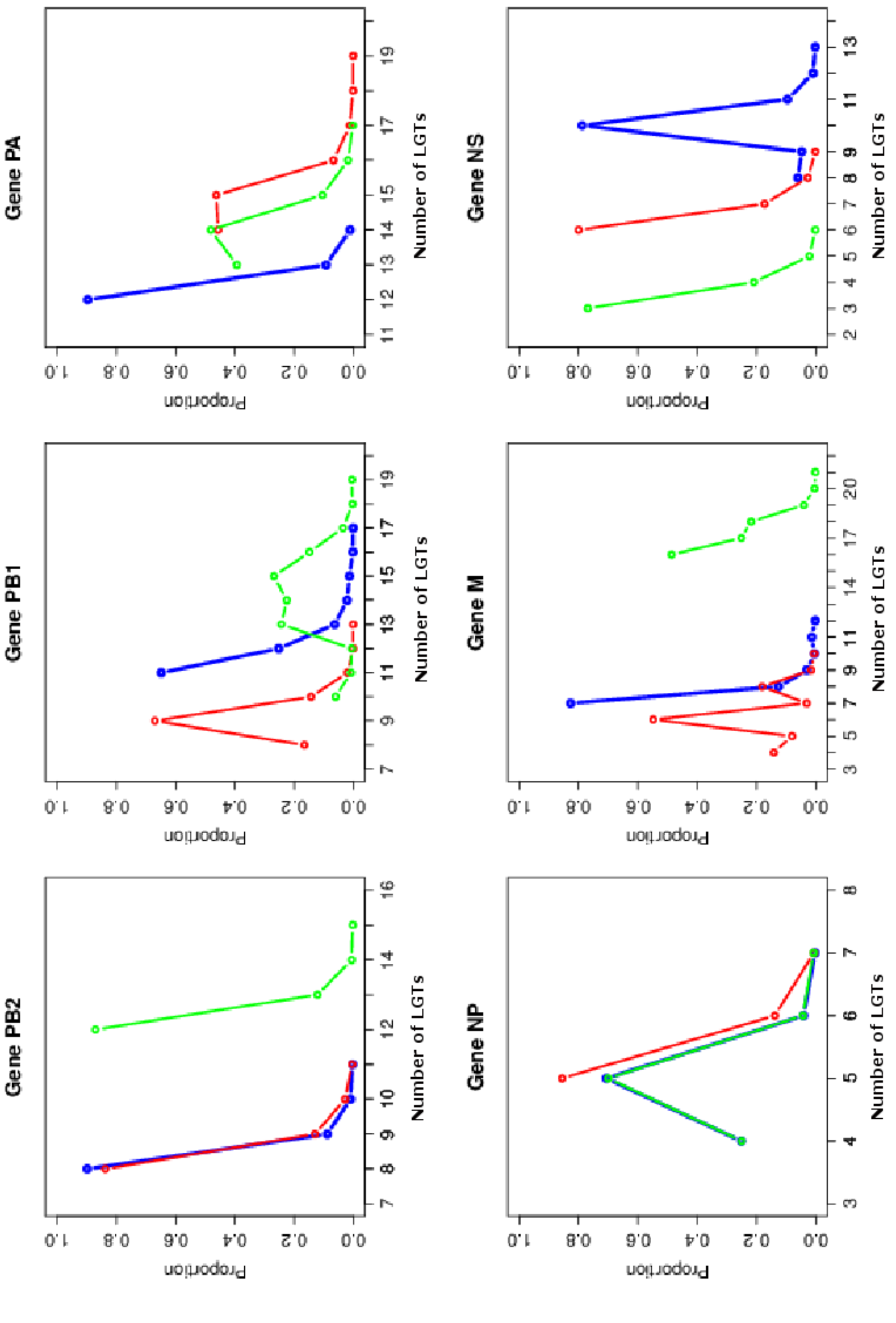
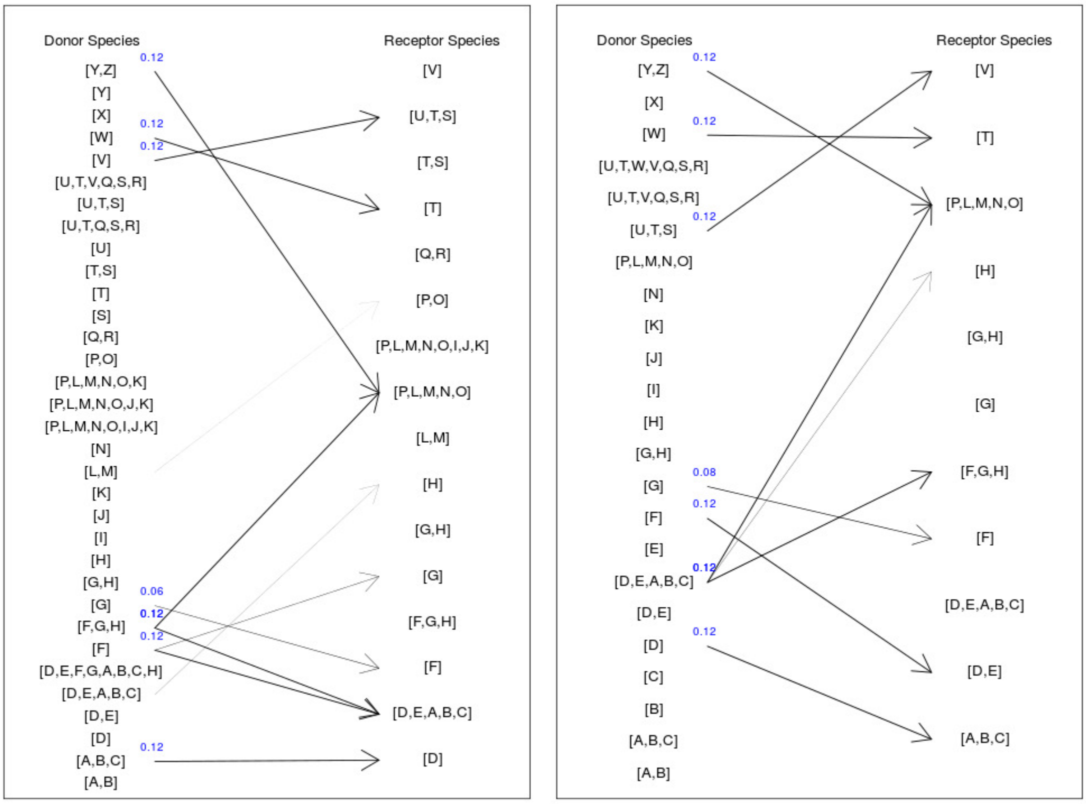


Figure 6.5: Number of LGTs for all genes and MCMC runs.

not involved in the outbreak, will be omitted from this analysis.



Gene PB2 - Red Chain

Gene PB2 - Blue Chain

Figure 6.6: Representation of the LGTs present in the posterior distribution for gene PB2.

Similarly to the results obtained by Smith *et al.* (2009), and common to both (red and blue) LGT histories is a gene transfer from the ancestor of the avian clade [Y,Z] to the clade containing the triple-reassortant swine and the 2009 human outbreak [L,M,N,O,P]. However, our results also suggest that this gene has genetic information from the human lineage. Both LGT histories agree that gene PB2 might have been also transferred from the human lineage, although the blue chain suggests a transfer from the H1N1 human lineage [F,G,H], while the red suggests that it has its origin in the clade containing the H2N2 and H3N2 lineages [A,B,C,D,E]. One possible explanation might be the presence of a recombination event in addition to reassortment. If more than one subtype of influenza virus replicates simultaneously in the cell nucleus, the genetic material of different subtypes of influenza viruses can recombine. As a result, each gene segment might contain genetic material from other subtypes. These results might suggest the sensitivity of our model to detect lateral transfer of genetic material on a minor scale when compared to reassortment events.

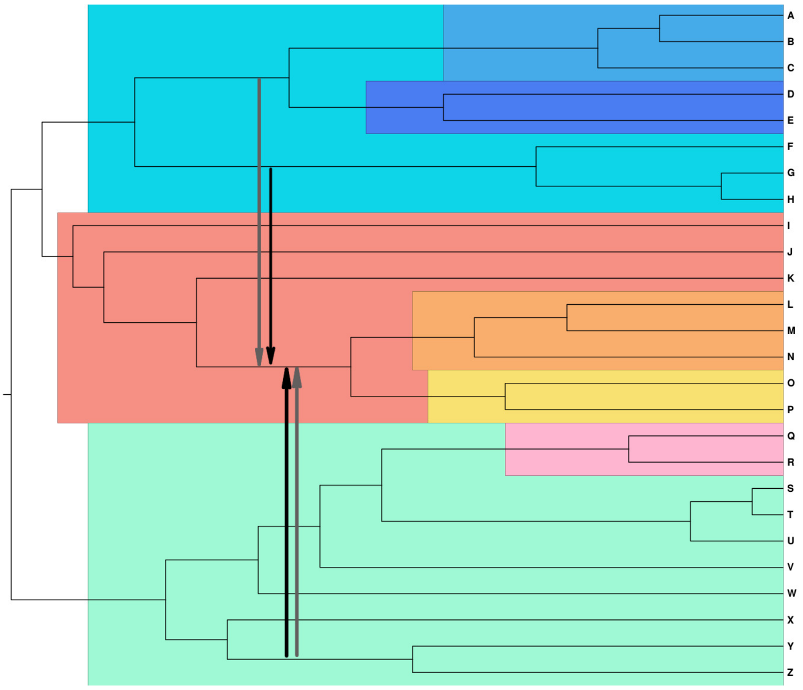
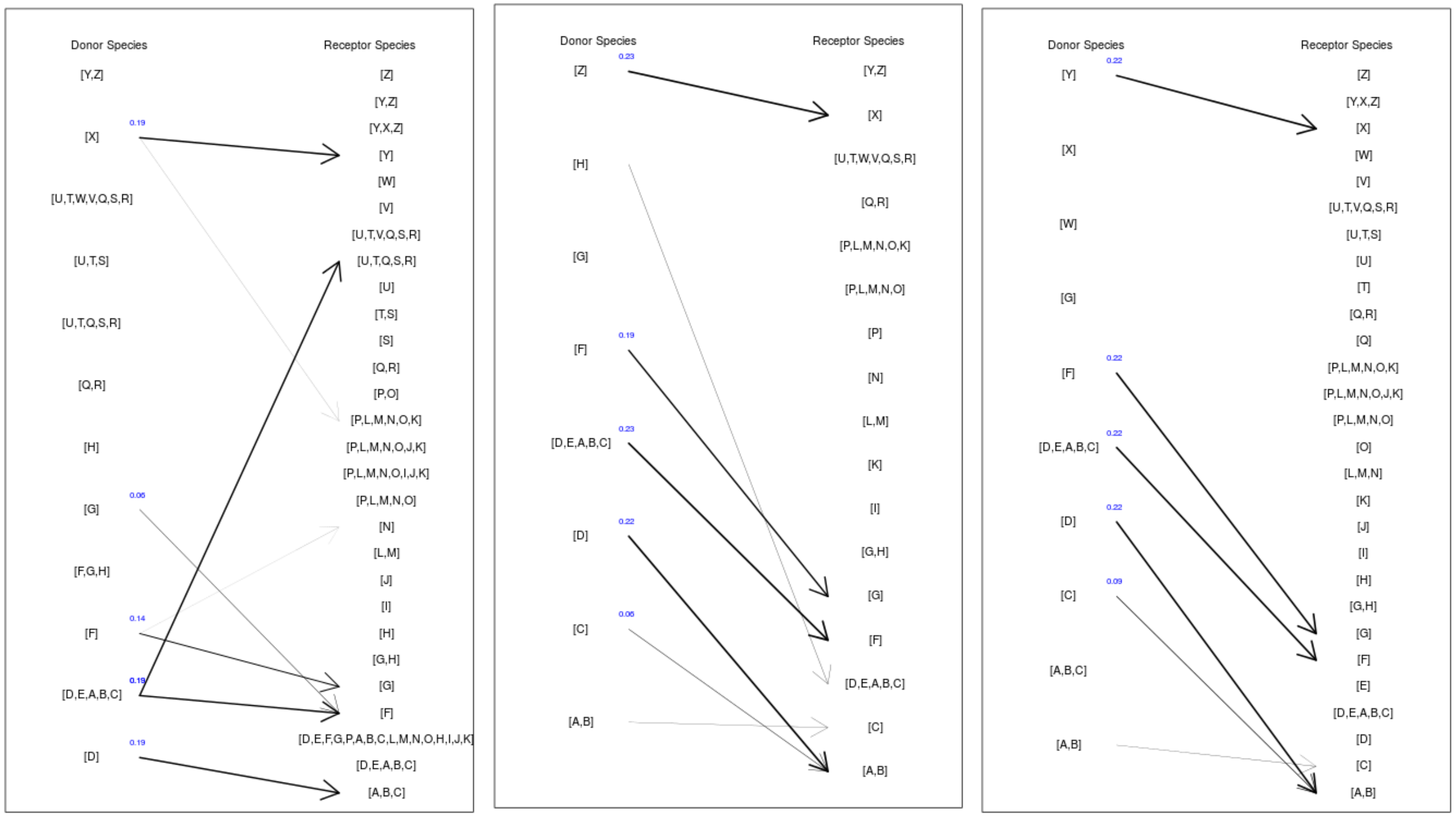


Figure 6.7: 2009 human S-OIV related LGTs for gene PB2 are represented in black and grey arrows for the red and blue chain, respectively.

In relation to gene NP, Figure 6.8 shows the LGT Biplots for all three chains. All three LGT histories are very similar and around 80% of the posterior probability corresponds to four LGTs in each chain. The discrepancies between the chains show a clear uncertainty on the relationship between avian species X, Y and Z, but the LGT directions suggested by the other three LGTs are similar in all three chains. Note that most LGTs occur within very closely related species and not between the major lineages, in special the one including the 2009 S-IOV viruses. This shows no evidence that the NP gene present in the 2009 S-OIV outbreak virus was laterally transferred. This result agrees with Smith *et al.* (2009) that the NP virus had its origin in the classical swine lineage.



Gene NP - Red Chain

Gene NP - Blue Chain

Gene NP - Green Chain

Figure 6.8: Representation of the LGTs present in the posterior distribution for gene NP.

7

Conclusion and Future Work

Different genes suggesting different evolutionary histories for the same group of organisms undermines the reconstruction of the underlying species tree from a set of gene trees. Beyond tree construction, the correct identification of genes that have undergone LGT is also an important biological problem, since it sheds light on the molecular pathways in which the genes play a role.

A principled statistical model for detecting LGT events in gene trees involves a combination of several hierarchical levels and combined simultaneous inference of gene trees and their relationship to an underlying species tree. Suchard (2005) developed a Bayesian approach to joint estimation of gene trees and an underlying species tree in the presence of LGT.

We have presented a method to reconstruct gene-related LGT histories which draws heavily on Suchard's work but takes a more biologically realistic approach by assuming an ordered rooted species tree, species contemporaneity on LGT events and site evolution rate heterogeneity. An extended version of the topological SPR operation (xSPR) was also introduced and, to enable inference using a Bayesian framework, MCMC proposals were developed for ordering a phylogenetic tree as well as a joint proposal for LGT history, LGT distance and gene trees.

A novel graphical representation, LGT Biplot, was also introduced as a useful and easily understandable way to visualise the gene transfer history.

Using simulated data under the conditions assumed by the model, we show that for all genes, all chains converged to the same posterior distribution (with strong support to the true value in all cases) for the substitution model parameters, branch lengths and topologies. In relation to LGT related parameters, i.e. number of LGTs and LGT history, the complexity of the parameter space, more specifically the fact that different LGT paths can result in the same gene tree topology given the species tree, often makes the transition between modes difficult. Several approaches were attempted to improve the mixing, such as different proposals for the LGT history, which proved to decrease significantly the percentage of chains that converged to LGT histories different from the true ones, on our simulated data. The good convergence of the posterior distributions for the HKY85+Γ substitution model parameters would encourage the use of more complex models, perhaps even the GTR+Γ.

Real data analysis was also performed in order to infer the 2009 S-OIV reassortment events by using our model and the data used in (Smith *et al.*, 2009), and compare our results with the ones obtained by the article. The data comprised multiple sequence alignments for 6 *Influenza* virus genes and 26 taxa. Although the MCMC runs did not reach overall convergence, for genes PB2 and NP some chains converged to a posterior distribution with a common topology and number of LGTs, as well as similar LGT histories. The LGT history recovered for both genes did correspond at some level to the conclusions reached by Smith *et al.* (2009) but our model seems to show additional sensitivity by detecting not only the reassortment events, but also lateral transfer of genetic material on a minor scale such as recombination events. As future work we could test for this using the multiple sequence alignments within a sequence-level analysis.

7.1 Future Work

Bridge proposal to improve mixing. It would be useful to have a proposal which would condition on S_o and the unrooted gene tree topology to give a sequence of xSPRs linking the two. Of course, shorter xSPR paths would be more likely, so the proposal should favour those. This could be achieved by making use of algorithms which compute the shortest xSPR path. For example: (i) compute shortest path between S_o and τ_g; (ii) with 90% probability do the first xSPR on this path, or with 10% probability an xSPR chosen uniformly at random; take the resulting stumpy tree and repeat the process. Nevertheless the shortest xSPR path is NP hard (it is as hard as the hardest problems in non-deterministic polynomial-time problems) to compute, though good heuristics exist, and it is not clear how to adapt the existing algorithms to get the shortest xSPR path. A better alternative would be to, when linking S_o to τ_g, propose xSPRs by weighting according to some score computed to reflect the similarity of each tree on the chain of xSPRs to the destination. For example, score according to the size of shared sub-trees. As disrupting shared sub-trees leads to unnecessarily large xSPR paths, by weighting we would avoid such xSPRs and obtain shorter paths.

Moves on *S*. Our model currently assumes a known species tree S which is biologically unrealistic in many analyses. In order to infer S it would be useful to find topological operations (NNI, xSPR) on S which are compatible with all the xSPR histories e.g. which affect part of S on which no LGTs occur. Nevertheless this is extremely restrictive and would lead to very poor mixing on S. Another idea is to use the bridge proposal described before: fix all the gene tree topologies, propose a new S via an NNI or xSPR, and then link the proposed S to the gene trees using the bridge proposal. This is another reason why developing a topological bridge would be very useful.

Allowing missing genes. Our current model requires that all genes in an analysis are present in every taxon. The model can be extended however to include a taxon with a missing gene by modelling the gene presence/absence on each edge of S_o using a birth-death model. Not only would this enable analyses of gene alignments with different number of taxa, it would also widen the type of LGT allowed to include 'copy and acquire' LGTs in addition to 'copy and overwrite'.

Cospeciation/Host-parasite model. Hosts and their associated parasites often exhibit a pattern of concordant phylogeny. Their phylogenies are largely congruent if the parasite tree is superimposed on the host tree, but not identical which might indicate that the phylogenies of hosts and parasites are not independent from one another (Huelsenbeck *et al.*, 1999). This might imply that some degree of host switching by the parasites has occurred. Our model can be adapted to infer the host gene tree topology (assumed to be the same as the host species tree), plus the sequence of SPRs corresponding to parasites switching host. For that we take a host gene and a parasite gene and fix $k = 0$ for the host gene, so that S_o has the same topology as the host phylogeny, and run the inference algorithm for the parasite genes.

Improving computational performance. Optimisation of the serial code and use of parallel computing techniques, by taking advantage of multicore processors, computer clusters, and GPUs, might potentially improve the performance of the current software with the purpose of releasing it to the scientific community.

Appendix A

Probabilistic Distributions

Dirichlet Distribution

The Dirichlet distribution, often denoted $\mathrm{Dir}(\boldsymbol{\nu})$, is a family of continuous multivariate probability distributions parametrized by a vector $\boldsymbol{\nu}$ of positive real numbers and is used for quantities describing proportions of a whole, so called *simplex* parameters. The Beta distribution, denoted $Beta(\nu_1, \nu_2)$ is equivalent to a Dirichlet that describes the probability on only two proportions, which are associated with the weight parameters $\nu_1 > 0$ and $\nu_2 > 0$. A Dirichlet distribution of dimension $K \geq 2$ with parameters $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$, where $\nu_k > 0$ for $k = 1, \ldots K$, has density

$$p\left(x_1, \ldots, x_{K-1} | \nu_1, \ldots, \nu_K\right) = \frac{1}{\mathrm{B}(\boldsymbol{\nu})} \prod_{i=1}^{K} x_i^{\nu_i - 1}, \qquad x_i \in (0,1) \tag{A.1}$$

on the open $(K-1)-$dimensional simplex defined by

$$x_1, \ldots, x_{K-1} > 0 \tag{A.2}$$

$$x_1 + \ldots + x_{K-1} < 1 \tag{A.3}$$

$$x_K = 1 - x_1 - \ldots - x_{K-1} \tag{A.4}$$

and zero elsewhere. The normalizing constant is the multinomial Beta function, which can be expressed in terms of the gamma function as

$$\mathrm{B}(\boldsymbol{\nu}) = \frac{\prod_{i=1}^{K} \mathit{\Gamma}(\nu_i)}{\mathit{\Gamma}\left(\sum_{i=1}^{K} \nu_i\right)}. \tag{A.5}$$

Let $V = \sum_{i=1}^{K} \nu_i$. Then X_i has mean $\frac{\nu_i}{V}$, variance $\frac{\nu_i(V - \nu_i)}{V^2(V+1)}$ and covariance $-\frac{\nu_i \nu_j}{V^2(V+1)}$. Examples include the stationary state frequencies that appear in the instantaneous rate matrix of the substitution model.

Exponential Distribution

The exponential distribution is a continuous distribution and has density function

$$p(x|\lambda) = \lambda e^{-\lambda x}, \qquad x > 0, \tag{A.6}$$

where λ is known as the rate parameter, for $\lambda > 0$. The mean and variance of an exponential distribution are $1/\lambda$ and $1/\lambda^2$, respectively. An exponential distribution is denoted by $Exp(\lambda)$.

Shifted Exponential Distribution

Sometimes it is useful to shift the exponential distribution away from zero. We define the shifted exponential distribution, $\mathrm{SExp}(\lambda, L)$, with density function

$$p(x|\lambda) = \lambda e^{\lambda(x-L)}, \qquad x \geq L, \tag{A.7}$$

where λ is the rate parameter, for $\lambda > 0$. The mean and variance of a shifted exponential distribution are $L + 1/\lambda$ and $1/\lambda^2$, respectively.

Gamma Distribution

The Gamma distribution, Gamma(α, β), is a two-parameter family of continuous probability distributions. Although there are three different parametrizations, for the purpose of this thesis, we assume a shape parameter α and rate parameter β where both parameters are positive real numbers. The Gamma(α, β) distribution has density function

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \qquad x > 0, \quad \alpha, \beta > 0, \tag{A.8}$$

with mean $\dfrac{\alpha}{\beta}$ and variance $\dfrac{\alpha}{\beta^2}$.

Geometric Distribution

The Geometric distribution is discrete and can have two different parametrisations depending on whether our interest is in modelling the number of trials until the first success (for $x = 1, 2, \ldots$) or the number of failures until the first success (for $x = 0, 1, \ldots$). Both parametrisations assume a known success probability θ, $0 \leq \theta \leq 1$ for each trial, and that the outcomes of the trials are independent. For the purpose of our model we give emphasis to the latter which has the probability function

$$p(x|\theta) = \theta(1-\theta)^x, \qquad x = 0, 1, \ldots, \tag{A.9}$$

with mean $(1-\theta)/\theta$ and variance $(1-\theta)/\theta^2$. This distribution can be abbreviated as Geom(θ).

Truncated Geometric Distribution

A random variable X has a truncated geometric distribution TGeom(θ,N), with parameters p and N, when it has probability function

$$p(x|\theta) = \frac{\theta(1-\theta)^x}{1-(1-\theta)^N}, \qquad x = 0, 1, \ldots, N. \tag{A.10}$$

Assuming that $q = 1 - \theta$, this distribution has mean

$$\frac{1 - q^N + q - Nq^N\theta}{(1-q^N)\theta} \tag{A.11}$$

and variance

$$\frac{(1+q^{2N})q - q^N(1+q^2)N^2 + q^{N+1}(N^2-1)}{\theta^2(1-q^N)^2} \tag{A.12}$$

(Olatayo, 2014).

Multinomial Distribution

For n independent trials, each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability, the multinomial distribution is a discrete distribution and gives the probability of any particular combination of numbers of successes for the various categories.

If $X_1, X_2, \ldots, X_n$ are mutually exclusive events with $Pr(X_1 = x_1) = \theta_1, \ldots,$ $Pr(X_n = x_n) = \theta_n$, where x_i are non-negative integers such that $\sum_{i=1}^{n} x_i = N$, and θ_i are constants with $\theta_i > 0$ and $\sum_{i=1}^{n} \theta_i = 1$, then the probability that X_1 occurs x_1 times, …, X_n occurs x_n times is given by

$$\frac{N!}{x_1! \ldots x_n!}\theta_1^{x_1} \ldots \theta_n^{x_n}. \tag{A.13}$$

The mean and variance of X_i are $\mu_i = n\theta_i$ and $\sigma_i^2 = N\theta_i(1-\theta_i)$. The covariance of X_i and X_j is $\sigma_{ij} = -N\theta_i\theta_j$.

Normal Distribution

The Normal (or Gaussian) distribution is a very commonly occurring continuous probability distribution. It is symmetric about its mean, and is non-zero over the entire real line. The normal distribution density is

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty < x < \infty, \mu \in \mathbb{R}, \sigma > 0 \tag{A.14}$$

where μ is the mean of the distribution and σ^2 is its variance.

Lognormal Distribution

The lognormal distribution, $\mathrm{LN}(\mu, \sigma^2)$, is a continuous probability distribution of a random variable whose logarithm is normally distributed. Given a lognormally distributed random variable X and two parameters μ and σ that are, respectively, the mean and standard deviation of the variable's natural logarithm, its probability density function is

$$p(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right), \qquad x > 0, \mu \in \mathbb{R}, \sigma > 0. \tag{A.15}$$

This distribution has mean $e^{\mu+\frac{\sigma^2}{2}}$ and variance $e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$.

Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space, assuming that these events occur with a known average rate w, independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$, if, for $x = 0, 1, 2, \ldots$, the probability mass function (p.m.f.) of X is given by

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{A.16}$$

The expected value of a Poisson-distributed random variable is equal to λ and so is its variance. The distribution is often abbreviated Po(λ).

Truncated Poisson Distribution

A discrete random variable X has a truncated Poisson distribution, TPo(λ, N), with parameters $\lambda > 0$ and $N > 0$ if, for $x = 0, 1, 2, \ldots, N$, the probability mass function (p.m.f.) of X is given by

$$p(x|\lambda) = \left(\sum_{i=0}^{N} \frac{\lambda^i e^{-\lambda}}{i!} \right)^{-1} \frac{\lambda^x e^{-\lambda}}{x!}. \tag{A.17}$$

Uniform Distribution

The discrete Uniform distribution is a probability distribution in which each of its say n elements are equally likely, each one with probability $1/n$. If the elements are defined as $\{a, a+1, \ldots, b = a+n-1\}$ then the distribution has mean $(a+b)/n$ and variance $(n^2-1)/12$. Uniform distributions are often used to express the lack of prior information for parameters that have a uniform effect on the likelihood in the absence of data. In phylogenetics, the discrete uniform distribution is typically used in relation to the topology parameter.

On the other hand, the continuous Uniform distribution is the probability distribution of a random number selected from a continuous interval (a, b), with $b > a$.

The distribution is often abbreviated as $U(a, b)$ and has density

$$f(x|a,b) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq x \leq b, \\ \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \tag{A.18}$$

with mean $(a+b)/2$ and variance $(b-a)^2/12$.

Appendix B

MCMC Results for Branch Lengths

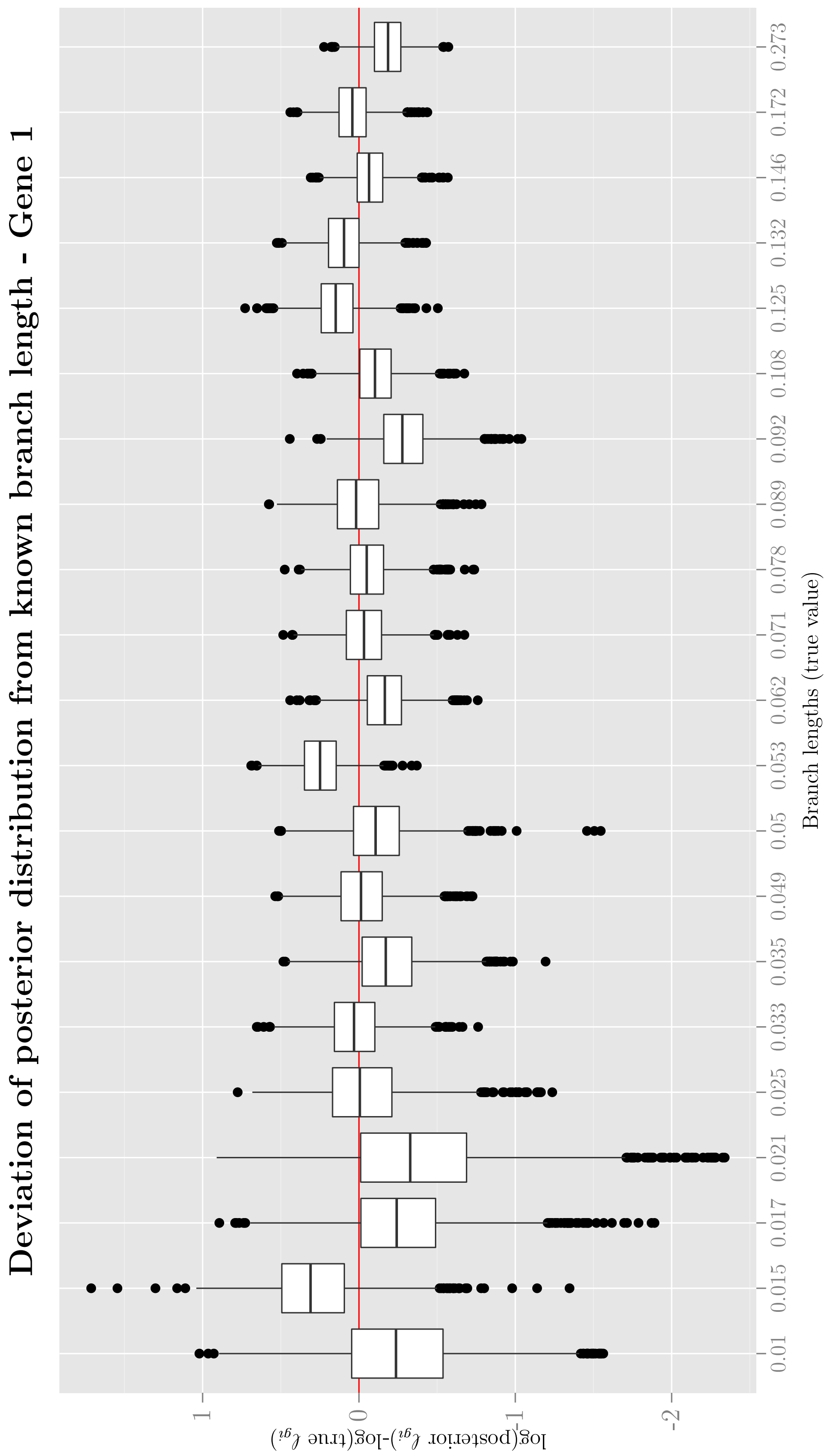


Figure B.1: Difference between the log of the posterior values and the log of the true value for branch lengths for gene 1.

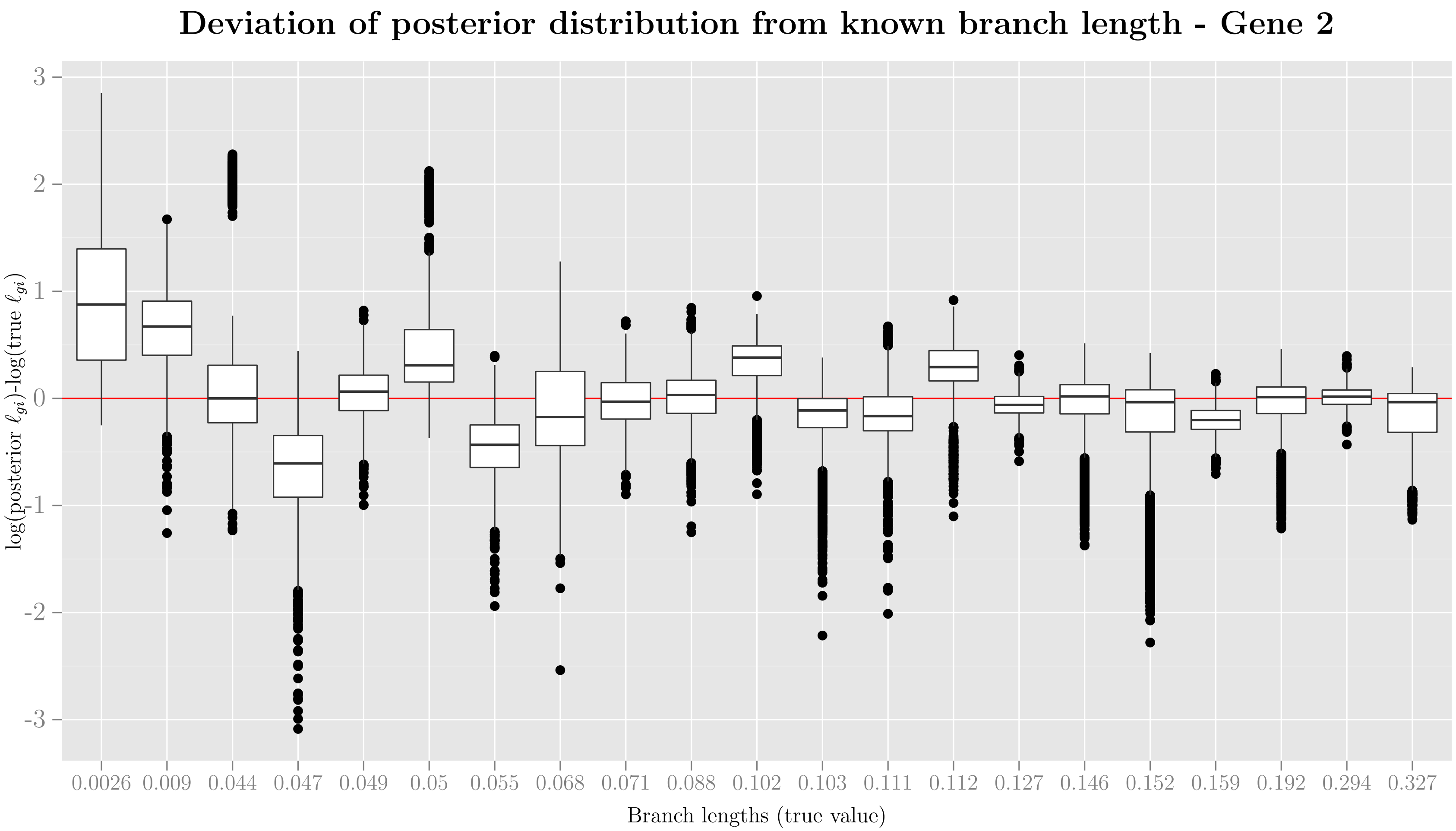


Figure B.2: Difference between the log of the posterior values and the log of the true value for branch lengths for gene 2.

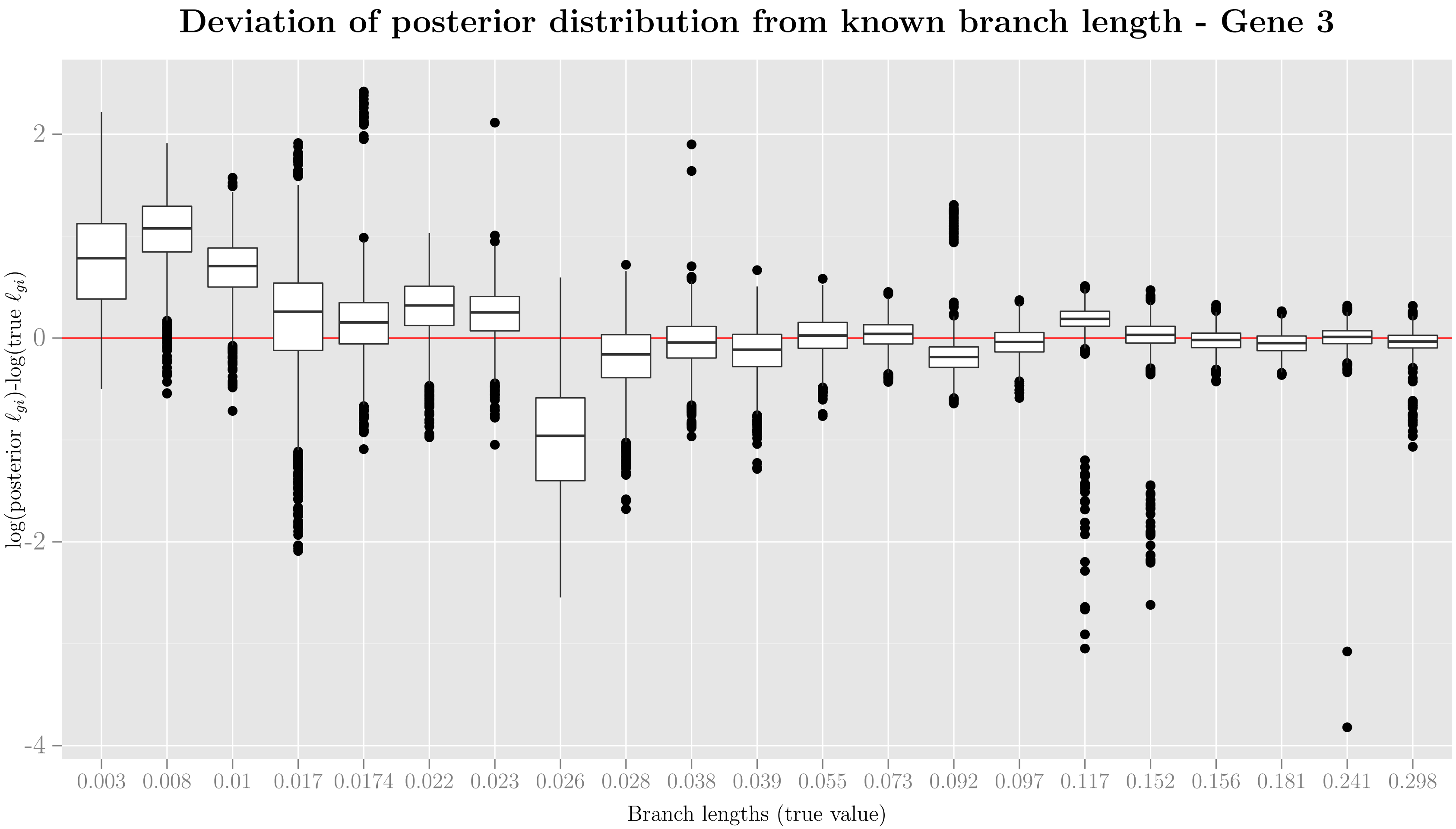


Figure B.3: Difference between the log of the posterior values and the log of the true value for branch lengths for gene 3.

Appendix C

Supplementary Results of Influenza Data Analysis

Log-scale		Natural-scale			
central tendencies		central tendencies		Measures of precision	
Param.	Mean (95% CI)	Param.	Mean (95% CI)	Param.	Mean (95% CI)
μ_ρ	2.42(1.96,3.05)	μ'_ρ	13.53(8.54,25.41)	$1/\sigma^2_\rho$	2.70(1.30,10)
μ_λ	3.18(2.73,3.66)	μ'_λ	28.36(18.08,45.83)	$1/\sigma^2_\lambda$	0.33(0.08,0.70)
		$\varPi_A$	0.34(0.30,0.39)	$N_{\varPi}$	72.43(28.10,119.36)
		$\varPi_G$	0.22(0.18,0.26)		
		$\varPi_C$	0.21(0.17,0.25)		
		$\varPi_T$	0.24(0.19,0.28)		

Table C.1: Across gene-level parameters for the Influenza data. For each parameter we have the posterior mean and 95% credible intervals.

Bibliography

Akaike, H. 1974 A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19(6)**, 716–723.

Allen, B. & Steel, M. 2001 Subtree transfer operations and their induced metrics on evolutionary trees. *Ann. Comb.* **5**, 1–13.

Andersson, J. 2005 Lateral gene transfer in eukaryotes. *Cell. Mol. LifeSci.* **62**, 1182–1197.

Bandelt, H., Macaulay, M. & Richards, M. 2000 Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol. Phyl. Evol.* **16**, 8–28.

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R. & Doolittle, W. 2005 Do orthologous gene phylogenies really support tree thinking? *BMC Evol. Biol.* **5**, 33.

Barker, D. 2004 LVB: parsimony and simulated annealing in the search for phylogenetic trees. *Bioinformatics* **20**, 274–275.

Beiko, R., Doolittle, W. & Charlebois, R. 2008 The impact of reticulate evolution on genome phylogeny. *Syst. Biol.* **57**, 844–856.

BEIKO, R., HARLOW, T. & RAGAN, M. 2005 Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci.* **102**, 14332–14337.

BOTO, L. 2010 Horizontal gene transfer in evolution: facts and challenges. *Proc. Biol. Sci.* **277(1683)**, 819–827.

BRADLEY, R., ROBERTS, A., SMOOT, M., JUVEKAR, S., DO, J., DEWEY, C., HOLMES, I. & PACHTER, L. 2009 Fast statistical alignment. *PLOS Computational Biology* **5**, e1000392.

BRIGULLA, M. & WACKERNAGEL, W. 2010 Molecular aspects of gene transfer and foreign DNA acquisition in prokaryotes with regards to safety issues. *Appl. Microbiol. Biotechnol.* **86**, 1027–1041.

BUCKLEY, T., ARENSBURGER, P., SIMON, C. & CHAMBERS, G. 2002 Combined data, Bayesian phylogenetics, and the origins of the New Zealand cicada genera. *Syst. Biol.* **51**, 4–18.

BULL, J., HUELSENBECK, J., CUNNINGHAM, C., SWOFFORD, D. & WADDELL, P. 1993 Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* **42**, 384–397.

CAVALLI-SFORZA, L. & EDWARDS, A. 1967 Phylogenetic analysis-models and estimation procedures. *Am. J. Hum. Genet.* **19(3Pt1)**, 233–257.

CLEMENT, M., POSADA, D. & CRANDALL, K. 2000 TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657–1660.

CRANSTON, K., HURWITZ, B., WARE, D., STEIN, L. & WING, R. 2009 Species trees from highly incongruent gene trees in rice. *Syst. Biol.* **58(5)**, 489–500.

DAGAN, T., ARTZY-RANDRUP, Y. & MARTIN, W. 2008 Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci.* **105**, 10039–10044.

DÁVALOS, L., CIRRANELLO, A., GEISLER, J. & SIMMONS, N. 2012 Understanding phylogenetic incongruence: lessons from phyllostomid bats. *Biol. Rev.* **87**, 991–1024.

DEGNAN, J. & ROSENBERG, N. 2006 Discordance of species trees with their most likely gene trees. *PLoS Genet.* **2(5)**, e68.

DENAMUR, E., LECOINTRE, G., DARLU, P., TENAILLON, O., ACQUAVIVA, C., SAYADA, C., SUNJEVARIC, I., ROTHSTEIN, R., ELION, J., TADDEI, F., RADMAN, M. & MATIC, I. 2000 Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103(5)**, 711–721.

DOOLITLE, W. & BAPTESTE, E. 2007 Pattern pluralism and the tree of life. *Proc. Natl. Acad. Sci.* **104**, 2043–2049.

DOOLITTLE, W. & LOGSDON, J. 1998 Archaeal genomics: do archaea have a mixed heritage? *Curr. Biol.* **8(6))**, R209–R211.

DRUMMOND, A. & RAMBAUT, A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214.

EDGAR, R. 2004 Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113.

ELLIOTT, T. 2014 Phylogenetic tree surgery 1. `http://telliott99.blogspot.co.uk/2010/11/phylogenetic-tree-surgery-1.html`, [Online; accessed 9-March-2014].

Excoffier, L., Smouse, P. & Quattro, J. 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Felsenstein, J. 1973a Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Am. J. Hum. Genet.* **25**, 471–492.

Felsenstein, J. 1973b Maximum likelihood estimation of evolutionary trees from continuous characters. *Systematic Zoology* **22(3)**, 240–249.

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

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

Felsenstein, J. 1989 Phylip – phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166.

Felsenstein, J. 2004 *Inferring Phylogenies*. Sunderland,MA: Sinauer Associates.

Fitch, W. 1971 Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**, 406–416.

Furuya, E. & Lowy, F. 2006 Antimicrobial-resistant bacteria in the community setting. *Nat. Rev. Microbiol.* **4**, 36–45.

Gadagkar, S., Rosenberg, M. & Kumar, S. 2005 Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *J. Exper. Zoology (Mol. Dev. Evol.)* **304B**, 64–74.

GALTIER, N. 2007 A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst. Biol.* **56(4)**, 633–642.

GALTIER, N. & DAUBIN, V. 2008 Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. Lond. (B)* **363**, 4023–4029.

GEMAN, S. & GEMAN, G. 1984 Stochastic relaxation, Gibbs distributions and the Bayes restoration of images. *IEEE Trans. Pattern Anal. Mach. Intel.* **6**, 721–741.

GEYER, C. 1991 *Markov chain Monte Carlo maximum likelihood. In* Computing Science and Statistics: Proc. 23rd Symp. Interface*, pp. 156-163*. Fairfax, VA: (ed. E.M.Keramidas) Interface Foundation.

GRAYBEAL, A. 1998 Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9–17.

HALLETT, M. & LAGERGREN, J. 2001 Efficient algorithms for lateral gene transfer. In *In Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB01)*, pp. 149 – 156.

HARTIGAN, J. 1973 Minimum evolution fits to a given tree. *Biometrics* **29**, 53–65.

HARVEY, P. & PAGEL, M. 1991 *Comparative method in evolutionary biology*. Oxford: Oxford University Press.

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.

HASTINGS, W. 1970 Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57(1)**, 97–109.

HEIN, J. 1993 A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.* **36**, 396–405.

HICKEY, G., DEHNE, F., RAU-CHAPLIN, A. & BLOUIN, C. 2008 SPR distance computation for unrooted trees. *Genetics* **4**, 17–27.

HIGGINS, D., BLEASBY, S. & FUCHS, R. 1992 Clustal v: improved software for multiple sequence alignment. *Comp. Appl. BioSc.* **8**, 189–191.

HOLDER, M. & LEWIS, P. 2003 Phylogeny estimation: traditional and Bayesian approach. *Science* **4**, 275–284.

HUELSENBECK, J., RANNALA, B. & LARGET, B. 1999 A Bayesian framework for the analysis of cospeciation. *Evolution* **54(2)**, 352–364.

HUELSENBECK, J., RONQUIST, F. & TESLENKO, M. 2012 Command reference for MrBayes 3.2. http://mrbayes.sourceforge.net/commref_mb3.2.pdf.

HUELSENBECK, J. P. & RONQUIST, F. 2001 MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.

HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R. & BOLLBACK, J. P. 2001 Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314.

HUSON, D. & BRYANT, D. 2006 Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23(2)**, 254–267.

HUSON, D., RUPP, R. & SCORNAVACCA, C. 2011 Phylogenetic networks: concepts, algorithms and applications. *Bioinformatics* **21(2)**, ii159–ii165.

IERSEL, L. & MOULTON, V. 2014 Trinets encode tree-child and level-2 phylogenetic networks. *J. Math. Biol.* **68(7)**, 1707–1729.

JUKES, T. & CANTOR, C. 1969 *Evolution of Protein Molecules*. New York: Academic Press.

Kaminski, L., Lurie-Weinberger, M., Allers, T., Gophna, U. & Eichler, J. 2013 Phylogenetic- and genome-derived insight into the evolution of N-glycosylation in archaea. *Mol. Phyl. Evol.* **68(2)**, 327–339.

Katoh, K., Kuma, K., Toh, H. & Miyata, T. 2005 MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518.

Keeling, P. & Palmer, J. 2008 Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9**, 605–618.

Kimura, M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16(2)**, 111–120.

Kingsford, C., Nagarajan, N. & Salzberg, S. 2009 2009 swine-origin influenza A (H1N1) resembles previous influenza isolates. *PLOS One* **4(7)**, e6402.

Kirkpatrick, S., Gelatt, C. & Vecchi, M. 1983 Optimization by simulated annealing. *Science* **220**, 4598.

Kluge, A. & Wolf, A. 1993 Cladistics: what's in a word. *Cladistics* **9**, 183–199.

Ku, C., Lu, W. & Kuo, C. 2013 Horizontal transfer of potential mobile units in phytoplasmas. *Mob. Genet. Elements* **3(5)**, e26145.

Kubatko, L., Salter, L. & Degnan, J. 2007 Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* **56(1)**, 17–24.

Kurland, C., Canback, B. & O.B.Berg 2003 Horizontal gene transfer: a critical view. *Proc. Natl. Acad. Sci.* **100**, 9658–9662.

Lartillot, N. & Philippe, H. 2004 A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21(6)**, 1095–1109.

Lederberg, J. & Tatum, E. 1946 Gene recombination in escherichia coli. *Nature* **158**, 558.

Lewis, P. 1998 A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.* **15(3)**, 277–283.

Li, S., Pearl, D. & Doss, H. 2000 Phylogenetic tree reconstruction using Markov Chain Monte Carlo. *J. Amer. Statist. Assoc.* **95(450)**, 493–508.

Lin, C., Bourque, G. & Tan, P. 2008 A comparative synteny map of Burkholderia species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. *Mol. Biol. Evol.* **25(3)**, 549–558.

Linder, C., Nakhleh, L. & Warnow, T. 2004 Network (reticulate) evolution: biology, models, and algorithms. In *In The Ninth Pacific Symposium on Biocomputing (PSB).*

Linz, S., Radtke, A. & von Haeseler, A. 2007 A likelihood framework to measure horizontal gene transfer. *Mol. Biol. Evol.* **24(6)**, 1312–1319.

Liu, L. & Pearl, D. 2007 Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* **56**, 504–514.

Löytynoja, A. & Goldman, N. 2008 Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632–1635.

Maddison, W. 1997 Gene trees in species trees. *Syst. Biol.* **46**, 523–536.

MAKARENKOV, V. 2001 T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics* **17(7)**, 664–668.

MARTYN, I. & STEEL, M. 2012 The impact and interplay of long and short branches on phylogenetic information content. *J. Theor. Biol.* **314**, 157–163.

MAU, B. & NEWTON, M. 1997 Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* **6**, 122–131.

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. & TELLER, E. 1953 Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21(6)**, 1087–1092.

MIRKIN, B., FENNER, T., GALPERIN, M. & KOONIN, E. 2003 Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**, 2.

MORGENSTERN, B., RINNER, O., ABDEDDAIM, S., HAASE, D., MAYER, K., DRESS, A. & MEWES, H. 2002 Exon discovery by genomic sequence alignment. *Bioinformatics* **18**, 777–787.

MURTAGH, F. 1984 Complexities of hierarchic clustering algorithms: the state of the art. *Computational Statistics Quaterly* **1**, 101–113.

NEUMANN, G., NODA, T. & KAWAOKA, Y. 2009 Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature* **459(7249)**, 931–999.

NIKOLAIDIS, N., DORAN, N. & COSGROVE, D. 2013 Plant expansions in bacteria and fungi: Evolution by horizontal gene transfer and independent domain fusion. *Mol. Biol. Evol.* **31(2)**, 376–386.

NOTREDAME, C., HIGGINS, D. & HERINGA, J. 2000 T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.

OCHMAN, H., LAWRENCE, J. & GROISMAN, E. 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* **405(6784)**, 299–304.

OLATAYO, T. 2014 Truncated geometric bootstrap method for time series stationary process. *Applied Mathematics* **5**, 2057–2061.

PAGE, R. 2000 Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Mol. Phyl. Evol.* **14**, 89–106.

PATTERSON, C. 1988 Homology in classical and molecular biology. *Mol. Biol. Evol.* **5**, 603–625.

POSADA, D. & WIUF, C. 2003 Simulating haplotype blocks in the human genome. *Bioinformatics* **19(2)**, 289–290.

RACANIELLO, V. 2013 Reassortment of the influenza virus genome. `http://www.virology.ws/2009/06/29/reassortment-of-the-influenza-virus-genome/`, [Online; accessed 30-November-2013].

RANNALA, B. & YANG, Z. 1996 Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304–311.

RANWEZ, V., HARISPE, S., DELSUC, F. & DOUZERY, E. 2011 MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLOS One* **6(9)**, e22594.

ROBERTS, G., GELMAN, A. & GILKS, W. 1997 Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.* **7**, 110–120.

Roberts, G. & Rosenthal, J. 1998 Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc. B* **60**, 255–268.

Roberts, G. & Rosenthal, J. 2001 Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.* **16**, 351–367.

Robinson, K., Sieber, K. & Hotopp, J. 2013 A review of bacteria-animal lateral gene transfer may inform our understanding of diseases like cancer. *PLoS Genet.* **9(10)**, e1003877.

Roch, S. & Snir, S. 2013 Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis **20(2)**, 93–112.

Rogers, J. & Swofford, D. 1999 Multiple local maxima for likelihoods of phylogenetic trees: A simulation study. *Mol. Biol. Evol.* **16(8)**, 1079–1085.

Ronquist, F., Mark, P. & Huelsenbeck, J. 2009 Bayesian phylogenetic analysis using MrBayes. In *The Phylogenetic Handbook: a practical approach to phylogenetic analysis and hypothesis testing* (ed. P. Lemey, M. Salemi & A. Vandamme). Cambridge: Cambridge University Press.

Rosenberg, N., Pritchard, J., Weber, J., Cann, H., Kidd, K., Zhivotovsky, L. & Feldman, M. 2002 Genetic structure of human populations. *Science* **298**, 2381–2385.

Rusin, L., Lyubetskaya, E., Gorbunov, K. & Lyubetsky, V. 2014 Reconciliation of gene and species trees. *BioMed Research International* **2014**, Article ID 642089, 22 pages.

Saitou, N. & Nei, M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4(4)**, 406–425.

SHI, Y., WU, Y., QI, J. & GAO, G. 2014 Enabling the 'host jump': structural determinants of receptor-binding specificity in influenza A viruses. *Nat. Rev. Microbiol.* **12(12)**, 822–831.

SLOWINSKI, J. & PAGE, R. 1999 How should species phylogenies be inferred from sequence data? *Syst. Biol.* **48**, 814–825.

SMITH, G., VIJAYKRISHNA, D., BAHL, J., LYCETT, S., WOROBEY, M., PYBUS, O., MA, S., CHEUNG, C., RAGHWANI, J., BHATT, S., PEIRIS, J., GUAN, Y. & RAMBAUT, A. 2009 Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125.

SOKAL, R. & MICHENER, C. 1958 A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **38**, 1409–1438.

SONG, Y. 2006 Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees. *Ann. Comb.* **10**, 129–146.

SONG, Y. S. 2003 On the combinatorics of rooted binary phylogenetic trees. *Ann. Comb.* **7**, 365–379.

SONG, Y. S. & HEIN, J. 2003 Algorithms in bioinformatics (proceedings of wabi 2003). pp. 287–302. Springer-Verlag, Berlin.

STEEL, M. 1994 The maximum-likelihood point for a phylogenetic tree is not unique. *Syst. Biol.* **43**, 560–564.

STEEL, M., LINZ, S., HUSON, D. & SANDERSON, M. 2013 Identifying a species tree subject to random lateral gene transfer. *J. Theor. Biol.* **322**, 81–93.

STOCKER, B., ZINDER, N. & LEDERBERG, J. 1953 Transduction of flagellar characters in salmonella. *J. Gen. Microbiol.* **9(3)**, 410–433.

SUCHARD, M. 2005 Stochastic models for horizontal gene transfer: Taking a random walk through tree space. *Genetics* **170**, 419–431.

SUCHARD, M., KITCHEN, C., SINSHEIMER, J. & WEISS, R. 2003 Hierarchical phylogenetic models for analysing multipartite sequence data. *Syst. Biol.* **52(5)**, 649–664.

SUCHARD, M., WEISS, R. & SINSHEIMER, J. 2005 Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. *Biometrics* **61**, 665–673.

SUSKO, E., FIELD, C., BLOUIN, C. & ROGER, A. 2003 Estimation of rates-across-sites distributions in phylogenetic substitution models. *Syst. Biol.* **5**, 594–603.

SWOFFORD, D. 2003 *Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sunderland, Massachusetts: Sinauer Associates.

SZÖLLOSI, G., BOUSSAU, B., ABBY, S., TANNIER, E. & DAUBIN, V. 2012 Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci.* **109(43)**, 17513–17518.

TRIFONOV, V., KHIABANIAN, H. & RABADAN, R. 2009 Geographic dependence, surveillance and origins of the 2009 influenza A (H1N1) virus. *N. Engl. J. Med* **361(2)**, 115–118.

WISECAVER, J., BROSNAHAN, M. & HACKETT, J. 2013 Horizontal gene transfer is a significant driver of gene innovation in dinoflagellates. *GenomeBiol. Evol.* p. first published online 19 November 2013.

WONG, K., SUCHARD, M. & HUELSENBECK, J. 2008 Alignment uncertainty and genomic analysis. *Science* **319**, 473–476.

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. 1995 Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. *J. Mol. Evol.* **40**, 689–697.

Yang, Z. 1996 Maximum likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**, 587–596.

Yang, Z. 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comp. Appl. BioSc.* **13**, 555–556.

Yang, Z. 2006 *Computational Molecular Evolution*. Oxford: Oxford University Press.

Yang, Z. & Rannala, B. 1997 Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14**, 717–724.

Yue, J., Sun, G., Hu, X. & Huang, J. 2013 The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate Monosiga brevicollis. *BMC Genomics* **14(1)**, 729.

Zhang, K. & Jin, L. 2003 Haploblockfinder: Haplotype block analyses. *Bioinformatics* **19(10)**, 1300–1301.

Zinder, N. & Lederberg, J. 1952 Genetic exchange in salmonella. *J. Bacteriol.* **64**, 679–699.

Zuckerkandl, E. & Pauling, L. 1965 Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (ed. V. Bryson & H. Vogel). New York: Academic Press.