RNA PHYLOGENETIC INFERENCE WITH HETEROGENEOUS SUBSTITUTION MODELS

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2006

By
Vivek Gowri-Shankar
School of Computer Science

Contents

Abstract 8

Declaration 10

Copyright 11

Acknowledgements 12

1 Introduction 13

1.1 The research context 13

1.2 The research problem 14

1.2.1 Molecular phylogenetics with RNA genes 15

1.2.2 Inference methods and substitution models 18

1.3 Aim of the thesis 20

1.4 Thesis structure 22

2 Substitution models and the likelihood function 24

2.1 Introduction 24

2.2 Phylogenetic inference from aligned sequences 26

2.2.1 Phylogenies 26

2.2.2 Parsimony methods 27

2.2.3 Distance methods 28

2.3 Nucleotide substitution models 30

2.3.1 Markov models of nucleotide substitution 30

2.3.2 Standard nucleotide models 33

2.4 RNA genes and base-pair substitution models 35

2.4.1 RNA molecules and compensatory substitutions 35

2.4.2 RNA substitution models 38

2.5 Phylogenetic trees and branch lengths in the likelihood framework 41

2.6 The likelihood function and the pruning algorithm 44

2.6.1 The pruning algorithm: simple case 44

2.6.2 Combined models and mixture models 48

3 Bayesian Phylogenetics 51

3.1 Introduction 51

3.2 Markov chain Monte Carlo methods 53

3.3 Priors 58

3.4 Proposal distributions 62

3.4.1 Proposals for continuous parameters 64

3.4.2 Proposals in the space of phylogenies 66

3.5 Practical considerations and issues 71

4 Heterogeneity in time 74

4.1 Introduction 75

4.2 A substitution process for compositional heterogeneity 78

4.3 MCMC proposals 81

4.3.1 Proposals for continuous parameters 81

4.3.2 Proposals in the space of topologies 82

4.3.3 Swapping composition vectors 83

4.4 Split and merge moves and the reversible jump MCMC method . 85

4.4.1 Reversible jump MCMC computation 85

4.4.2 Dimension changing proposals 86

4.4.3 Split and merge moves for the phylogeny 87

4.4.4 Split and merge moves from the pool perspective 88

4.4.5 Computation of the acceptance rate 90

4.5 A hierarchical Bayesian model 94

4.5.1 Doubts over the uniform prior on the allocation vector . . 94

4.5.2 Changes introduced by the hierarchical model 96

4.5.3 Comparison of the two models 97

4.6 Bayesian simulations 98

4.7 Tree of Life 102

4.7.1 G+C content and thermophily 102

4.7.2 Phylogenetic reconstruction 106
4.7.3 Ancestral G+C content 110
4.7.4 Amount of heterogeneity 112
4.8 Conclusion 113

5 Heterogeneity across sites 116
5.1 Introduction 117
5.2 Modeling rate variation across sites 120
5.2.1 Invariant sites model 120
5.2.2 Discrete gamma model 121
5.2.3 Allowing for auto-correlation between adjacent sites 122
5.3 Variation of equilibrium frequencies across sites 123
5.3.1 Estimating the composition for each rate category 123
5.3.2 Robustness of the method and results on artificial data . . 125
5.3.3 Application with real RNA genes of 13 primates 128
5.3.4 Discussion 130
5.4 Effects on standard phylogenetic inference methods 131
5.4.1 Simulations 131
5.4.2 Discussion 133
5.5 Modeling compositional heterogeneity across sites 134
5.5.1 A Gaussian Process model 134
5.5.2 Test with simulated datasets 137
5.5.3 Application to the primate dataset 137
5.5.4 Discussion 140
5.6 Effects on time-heterogeneous methods 141
5.6.1 Simulations 142
5.6.2 A mesophilic LUCA? 143
5.6.3 Discussion 144
5.7 Conclusion 146

6 Conclusion 148
6.1 Summary of the thesis 148
6.2 Phylogenomics and the future of PHASE 150
6.3 Possible extensions and future work 151
6.3.1 Improving the MCMC algorithms 151
6.3.2 Evolutionary models 152

A Jacobian for the split/merge proposals of composition vectors 154

B The normalizing factor Z used for the GP prior 157

List of Figures

1.1 Monophyletischer Stammbaum der Organismen 15
1.2 The Phylogenetic Tree of Life (current view) 16

2.1 Nucleotide alignment 26
2.2 Phylogenetic tree: definitions 27
2.3 Markov model of nucleotide substitution 30
2.4 Evolution of a nucleotide over time 32
2.5 The K80 and HKY85 substitution models 34
2.6 Schematic representation of the RNA secondary structure 36
2.7 ncRNAs secondary and tertiary structures 37
2.8 Compensatory substitution mechanism 38
2.9 The 7D substitution model 42
2.10 Unrooted and rooted trees 42
2.11 Ultrametric tree 44
2.12 Pruning algorithm 46

3.1 Integration with MCMC methods 55
3.2 Marginalization with MCMC 57
3.3 The coin tossing problem in a Bayesian perspective 59
3.4 Nearest Neighbor Interchange for unrooted trees 67
3.5 Nearest Neighbor Interchange for ultrametric trees 68
3.6 Subtree Pruning and Regrafting for unrooted trees 69
3.7 Subtree Pruning and Regrafting for ultrametric trees 70
3.8 Branch proposal for ultrametric trees 71

4.1 The compositional bias illustrated with a nuclear LSU dataset . . 77

4.2 The NNI with the nonhomogeneous substitution model 83
4.3 The SPR with the nonhomogeneous substitution model 84
4.4 Reallocation during a split/merge proposal 88
4.5 Prior on the number of composition vectors 91
4.6 MAP estimates for the number of composition vectors 98
4.7 BPP distributions for the number of composition vectors 99
4.8 Comparison of BPPs with frequentist probabilities 101
4.9 Optimal growth temperatures vs. rRNA G+C contents 105
4.10 Consensus tree with the bacterial rooting 108
4.11 Consensus tree with the eukaryotic rooting 109
4.12 Ancestral G:C content 111
4.13 Posterior probability densities for the ancestral G:C contents . . . 111
4.14 BPP distribution for the number of composition vectors 113
4.15 BPP distribution with different priors 114

5.1 The gamma model for across-sites rate variation 121
5.2 The tree that generated the synthetic datasets 126
5.3 Compositional variation across sites with simulated data 127
5.4 Compositional variation across sites in mitochondrial RNA genes . 129
5.5 Application of the GP prior model to synthetic data 138
5.6 Application of the GP prior model to real data 139
5.7 Frequency variation across sites in the Tree of Life dataset 145

Abstract

Current Maximum Likelihood and Bayesian phylogenetic methods are based on highly simplified probabilistic models of sequence evolution. To improve the accuracy of the reconstructed evolutionary trees and to increase our understanding of the evolutionary processes at the molecular level, it is important to introduce more biological realism into the underlying evolutionary models. Indeed, wrong evolutionary assumptions often introduce a significant bias which might lead to incorrect reconstructions and inconsistent results. The aim of the research presented in this thesis is to provide improved methods for RNA-based phylogenetics. To that end, the **PHASE** software package, which allows evolutionary tree reconstruction with specific models accounting for the conserved secondary structure of RNA genes, is extensively rewritten. New features and models are implemented.

Current methods that model the nucleotide substitution process assume homogeneity of nucleotide composition among different lineages. Yet there is strong evidence that nucleotide frequencies are varying along different lineages in nuclear and mitochondrial RNA genes. Failure to account for the heterogeneity of the evolutionary process over time can lead to the recovery of spurious phylogenies. Homogeneous methods tend to group together species which are related in terms of nucleotide composition rather than in terms of evolutionary history. Following earlier work from other researchers, a time-inhomogeneous substitution model using different evolutionary parameters on different branches of the tree is developed. Using a reversible jump Markov chain Monte Carlo technique, the model can statistically detect the amount of heterogeneity exhibited by the data without overfitting. The method is tested on both synthetic and empirical datasets.

A second strand of this research is concerned with the variation of nucleotide frequencies among sequence sites. Compositional variation in time has already been extensively studied but fewer studies have focused on the effects of compositional heterogeneity within genes. It is shown here that different sites in an alignment do not always share a unique compositional pattern. Examples are provided where nucleotide frequency trends are correlated with the site-specific rate of evolution in RNA genes. Spatial compositional heterogeneity is shown to affect the estimation of evolutionary parameters and a new model to account for compositional variation across sites is developed. A Gaussian process prior is used to allow for a smooth change in composition with evolutionary rate. The results suggest that this model can accurately capture the observed trends in present-day RNA sequences.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Acknowledgements

First and foremost, I would like to express my thanks and gratitude to my supervisor, Dr Magnus Rattray, for comments and guidance through all this research project. His sound advice and experience rescued me from difficult situations and sent me in the right direction more than once. I would like to thank my predecessor Howsun Jow who wrote the first version of PHASE and my initial collaborators Paul Higgs and Cendrine Hudelot. Their help was invaluable and provided me with a head start. I also wish to thank Bastien Guillard and Antoine Buxerolles who helped with software development and implemented protein and codon models during their MSc projects.

I am grateful to the entire AI group and my fellow Ph.D. students who shared an office and/or the coffee maker with me: Jason Fleischer, Susannah Lydon, Richard Craggs, Elon Correa, Rob Woolfsoon, Gilles Daniel, Xuejun Liu, Shenghui Wang, Hossein Sharif Paghaleh and Richard Pearson. Special thanks go to Gwenn Englebienne, Fabrice Caillette and Lilian Janin for correcting English style and grammar mistakes in this thesis and offering useful suggestions.

I also owe a lot to the first users of PHASE. It was very encouraging and stimulating to learn that the software was used by others and it really helped to keep me motivated. The feedback they provided was very helpful.

Finally, I would like to thank my parents and Jun for constant care and support.

Chapter 1

Introduction

1.1 The research context

The theory of evolution states that all organisms are related through a history of common descent. It is widely accepted that life on Earth diversified in a tree-like pattern and that all living species can ultimately be traced back to a single common ancestor at the root of the tree. The idea that species could change and evolve over time is more than one hundred years old (Lamarck, 1809; Darwin, 1859) but the Tree of Life is still not completely resolved today. Admittedly, huge progress has been made since the first published phylogenies (*e.g.*, Figure 1.1) but the current picture (Figure 1.2) is not yet set in stone and many branches and deep bifurcations are still unknown (*Science*, *300*, special issue). Accordingly, one aim of this thesis is to improve on the methods currently available to reconstruct phylogenetic trees. Another related aim is to study the differences between genes in an evolutionary perspective in order to improve our understanding of the processes of DNA sequence evolution.

In the field of Systematics, phylogenetic trees are naturally used to classify species according to their evolutionary relationships. Reconstructing evolutionary trees for the sake of knowledge alone is certainly a worthwhile endeavour. However, phylogenies also have more practical use and a wide range of applications in biology (Harvey et al., 1996). The evolutionary paradigm has been central to biology for over a century and Life Sciences acquire a whole new dimension from an evolutionary perspective. For instance, recovering the evolutionary history of genetic systems and metabolic pathways can help to understand their current roles and mechanisms of action. Phylogenies also have a notable impact in the field of epidemiology. They have been used to study HIV transmission (Ou et al., 1992; Sharp et al., 1995) and to identify sites under positive selection (Yang, 2001). Recently, they have also been used to characterise the SARS-associated virus as a new type of coronavirus (Marra et al., 2003). Evolutionary trees have also been proposed as strategic tools to measure biodiversity and draw up conservation priorities (Mace et al., 2003).

1.2 The research problem

The main research problem is easily stated: given a set of *homologous* characteristics (*i.e.*, having the same unique evolutionary origin) from a group of species or viruses, infer the tree that represents their evolutionary relationships. The first concern to reconstruct phylogenies is consequently the choice of data to perform the task. Some phylogenetic markers perform well for closely related species but their phylogenetic signal is quickly eroded by the passage of time. Others evolve too slowly to be of any use except to resolve deep bifurcations in early history inference.

Molecular markers have proven very useful to resolve phylogenetic trees in recent years and a second research problem appeared. DNA sequences can be used to reconstruct phylogenies but knowledge of the evolutionary history can, in turn, help us to understand how these DNA sequences are evolving. Methods developed in this thesis are designed to infer the evolutionary processes that shaped contemporary molecules simultaneously to the pattern of branching. Although the reconstruction of accurate phylogenies is an important issue, some emphasis is placed on the understanding of evolutionary mechanisms at the molecular level in this thesis.

Since the Earth preserved few traces of its past, these two related tasks are usually complicated by the fact that data from long-dead ancestral species are missing. The tree has to be reconstructed from contemporary data only. Defining a criterion to evaluate how good a phylogeny is, gauging how well it explains the contemporary observed characteristics and their relatedness, comparing the

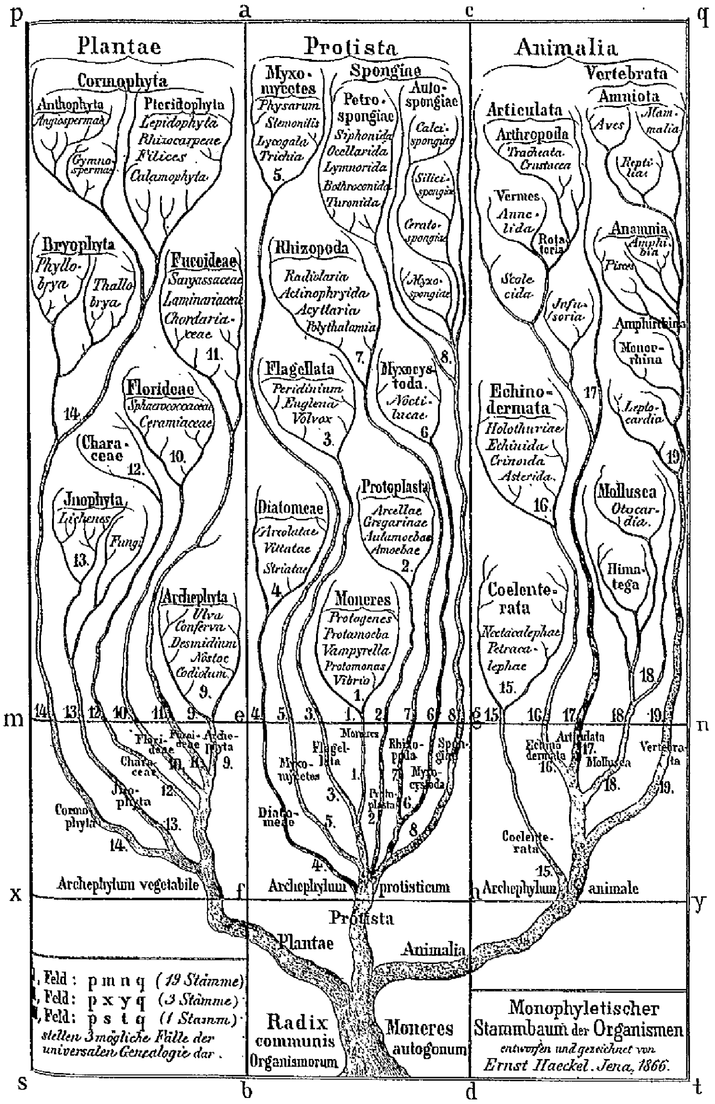


Figure 1.1: Monophyletischer Stammbaum der Organismen (Haeckel, 1866, II: plate I).

relative merits of alternative phylogenies, assessing the reliability of the results and complementing them with confidence values, are all underlying issues that have to be addressed in the process.

1.2.1 Molecular phylogenetics with RNA genes

Theoretically, any evolving characteristic can be used as a basis to reconstruct phylogenies. Phenotypic data, *i.e.*, the observable physical and behavioural traits

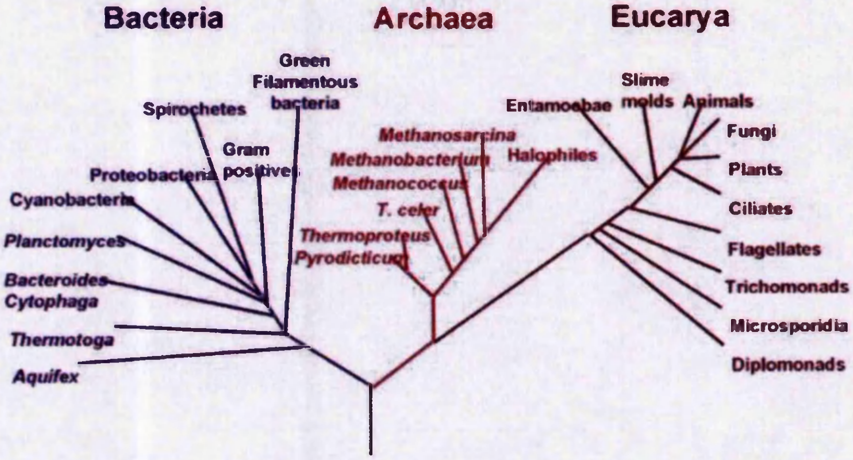


Figure 1.2: The phylogenetic Tree of Life — current view based on phylogenetic inference with the SSU rRNA gene (Woese and Fox, 1977).

of organisms, were traditionally used before DNA sequences became widely available. However, the phylogenetic signal is often distorted when using such data because evolutionary unrelated species often evolve similar characteristics independently when living in similar environments and confronted with the same evolutionary challenges. For instance, there is evidence that wings and eyes have been developed more than once because of this phenomenon of convergent evolution (Harvey and Pagel, 1991). The genetic information is the most basic data subject to evolutionary change. The rapid accumulation and distribution of macromolecular data — mainly DNA and protein sequences — greatly reduced the contribution of phenotypic data in this field of research and biomolecules became the most popular form of data in phylogenetic inference over the past 20 years.

The DNA molecules of an individual contain the genetic information that determines its growth and development. DNA chains are linear sequences of nucleotides (**A**denine, **C**ytosine, **G**uanine or **T**hymine) and specific regions of the DNA, called the genes, encode the information necessary to produce functional proteins. Over evolutionary time-spans, these genes change and the genetic sequences are altered. The aim of molecular phylogenetics is to infer the evolutionary relationships of a group of species using the similarities and differences between their genetic information. Genes evolve by accumulating changes. Intuitively, the number of differences between homologous genes taken from two different species should be correlated to the age of their common ancestor. The higher the number of differences, the larger the evolutionary distance between the two species. Opposedly, genes of closely related species, belonging to the same group, are expected to be more similar.

Most well characterised genes are protein-coding genes and are first *transcribed* into messenger RNAs (mRNAs) before being *translated* into functional proteins. Not all genes are protein-coding sequences. Some RNA molecules are not transporters of information but have a functional role in the organism. These specific RNAs are directly transcribed from the non-coding RNA genes (ncRNA genes). In this thesis the focus is on these functional RNA molecules. ncRNA genes from contemporary species are the data used to perform phylogenetic inference. The ncRNA genes investigated in this work are the genes coding for ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs). These RNA molecules play a crucial role during the translation of mRNA into proteins and are consequently ubiquitous in all known life forms.

The shape of an RNA structure often determines its function. The evolutionary forces acting to change RNA-coding sequences naturally have to thread around the structure of these molecules. In this thesis, the importance of bonding interactions between nucleotides during the evolutionary process is recognized and the phylogenetic inferences are based on methods that account for the secondary structure of the RNA products.

The Large Subunit (LSU) and Small Subunit (SSU) nuclear rRNA genes are slow-evolving genes that are traditionally used to reconstruct phylogenies. Since these rRNA genes have been sequenced for a large number of species, they are traditionally used as a reference to compare with alternative phylogenies based on different data. By far, nuclear rRNA genes have been the most widely used to reconstruct the deep branching patterns of the Tree of Life and they will be used for that purpose in this thesis.

Most eukaryotic cells contains mitochondria. Mitochondria are distinct isolated organelles which are involved in the respiration process and provide the cell with energy. These "cells within the cell" also have a genetic material which contains ncRNA genes. Although mitochondrial rRNAs have the same evolutionary origin as nuclear rRNAs, they are evolving much faster and are consequently useful to resolve close evolutionary relationships. Mitochondrial RNAs, *i.e.*, rRNAs and tRNAs, are used to produce mammalian phylogenies in this thesis.

Although the emphasis is on rRNAs and tRNAs, the methods developed in this thesis are by no mean not limited to these specific RNA molecules. They could also be used with RNA viruses or other small RNAs as long as a reasonably sized dataset can be built. Moreover, most techniques introduced in this thesis can easily be modified to be used in "traditional" molecular phylogenetics with DNA or protein sequences.

1.2.2 Inference methods and substitution models

A bewildering variety of inference methods are now available to reconstruct phylogenetic trees from molecular data. Modern approaches to phylogeny, pioneered by Felsenstein (1981), are based on probabilistic foundations and use the *likelihood* as their optimality criterion. Although other inference methods are briefly outlined at the beginning of chapter 2, likelihood-based methods are the main subject of this thesis.

Likelihood-based methods require the specification of an explicit substitution model. Simply put, this substitution model gives the probability that a nucleotide is substituted with another one between two time-points. This substitution model is ultimately used to compute the likelihood function, *i.e.*, the probability that a given evolutionary model (tree + substitution model) has generated the observed sequences. Likelihood-based methods have many advantages over the more traditional approaches. They permit the inference of phylogenetic trees using complex and realistic models of sequence evolution. Moreover, uncertainty about the process of substitution can be introduced in the model using "free" parameters (*e.g.*, the substitution rate from one nucleotide to another, the frequency of a specific nucleotide, etc) which can be estimated along with the pattern of descent during the inference process.

So far, the Maximum Likelihood (ML) method has been the main statistical approach in this field of research. The tree that yields the highest likelihood, over all possible patterns of branching, associated branch lengths and evolutionary model parameters, is considered to be the "best" tree. Standard optimization algorithms can be used to recover ML estimates for the branch lengths and the free parameters of the substitution model since these are continuous parameters.

However, finding the best phylogeny in the discreet topology space is much harder. Only an exhaustive search, possibly associated with branch and bound methods, is guaranteed to recover the optimal tree. The number of possible phylogenies rises as the factorial of the number of species. For as few as twelve species more than ten billion trees are possible (Felsenstein, 2004). Due to the intractable size of the tree topology space, various heuristic methods are currently used in order to drastically cut down the time needed to optimize each candidate tree (see, *e.g.*, Guindon and Gascuel, 2003) and to direct the search towards the optimal topology. As a result, these heuristic algorithms often find a near-optimal tree topology rather than the correct one.

Another issue with the ML method is that it singles out a specific tree and a unique set of ML estimates for the evolutionary parameters. This would not be such an issue if the assumed model of sequence evolution was correct and if sequences were long enough. Indeed, ML estimators are known to be consistent and efficient (Chang, 1996b; Rogers, 1997; Yang, 1997a), which means that the correct evolutionary tree will always be reconstructed given enough data (nevertheless, see Farris, 1999, for a contrary viewpoint). However, these two conditions are not met in practice and some uncertainty remains after an analysis. Methods have been developed to address these concerns. For example, the nonparametric bootstrap is often used to measure confidence in a reconstructed tree and to evaluate how reliable a ML estimate is (Felsenstein, 1985). The statistical foundations of the likelihood also allow the rigorous comparison of different substitution models (Goldman, 1993) and the evaluation of the relative merits of different phylogenies (Kishino and Hasegawa, 1989; Shimodaira and Hasegawa, 1999).

In this thesis, the emphasis is not on the ML approach but on the related Bayesian approach, which was introduced in this field of research by Mau (1996) and Li (1996) in their PhD thesis,Yang and Rannala (1997) and Larget and Simon (1999). The Bayesian approach combines the information contained in the data (using the likelihood function) with some prior information or belief to generate the *posterior probability distribution* of the parameters of interest. In the Bayesian framework, results are not point-estimates but probability distributions. Inferred evolutionary parameters and phylogenies automatically come with a measure of uncertainty. Results presented in this thesis, phylogenies and/or model parameter estimates, are consequently accompanied with posterior probabilities, descriptive statistics (usually mean posterior estimates) or 95% credibility intervals derived from these. These posteriors have the important advantage of being based on the integrated likelihood. The posterior probability of a topology is averaged over branch lengths and substitution parameters (sometimes called nuisance parameters) and, reciprocally, the credibility interval of a specific substitution parameter is integrated over the uncertainty in the topology.

In phylogenetic inference, the computation of posterior probabilities cannot be derived analytically but numerical Monte Carlo methods can be used to sample from the probability distribution. Markov chain Monte Carlo (MCMC) algorithms (Metropolis et al., 1953; Hastings, 1970) are the major computational methods used to approximate the posterior probability distribution. The computational efficiency of MCMC methods is responsible for the current popularity of the Bayesian approach in phylogenetic inference. Indeed, the surge of interest over Bayesian techniques is not really related to the philosophical debate opposing frequentists and Bayesians but is rather a practical choice. MCMC methods are fast and the Bayesian approach makes it possible to handle and evaluate complex evolutionary models. The Bayesian framework is consequently very attractive for researchers with a focus on the sequence evolution process rather than on the tree.

1.3 Aim of the thesis

In order to make substitution models both computationally and mathematically tractable, many simplifying assumptions are made. Most of the research presented in this thesis is actually concerned with relaxing the most unrealistic restrictions and proposing better substitution models for the evolution of RNA genes. The aim of the research is at least twofold. First, it is believed that more realistic models can recover evolutionary relationships more faithfully. Much effort is currently being put into the development of better evolutionary models for protein-coding genes (Thorne et al., 1996; Liò and Goldman, 2002). Since RNA-based phylogenetics also has a huge influence in the field of evolutionary history, it is only natural to design substitution models better suited to the evolution of RNA genes. Second, likelihood-based methods are not limited to the reconstruction of phylogenetic trees and can be used to infer the process of evolution simultaneously to the tree topology. A second aim of this research is consequently to improve our understanding of RNA evolution mechanisms by finding out which features of the models are the most important to explain the observed sequences and fit the data accurately.

In the early days of likelihood-based phylogenetic methods, it was traditionally assumed that:

1. the substitution process is Markovian and the future substitutions at a specific nucleotide position do not depend on past substitutions (*i.e.*, the process has no memory);
2. each site is independent and a substitution at a site does not affect the process at other sites;
3. the process is shared across sites and the same Markov model is used at each position;
4. the substitution process is homogeneous over time and over lineages and free parameters of the substitution process are constant;
5. the Markov process is at equilibrium and expected nucleotide frequencies are stationary over the tree, including the root (common ancestor) and the leaves (contemporary species).

As shall be shown in this thesis, these assumptions are unfortunately not very realistic for the considered RNA genes. Methods to relax all but the first constraint are reviewed and proposed. The substitution models considered in this work are all Markovian but note that this limitation was also discussed and addressed by other researchers (Benner et al., 1994; Crooks and Brenner, 2005).

New methods are based on base-pair substitution models that relax the second assumption and account for the evolutionary correlation between the two nucleotides of paired-sites in RNA helices (described in chapter 2). Jow (2003), implemented some of these models in a software package called **PHASE**. This phylogenetic inference package was rewritten and extended to implement the new methods presented in this thesis. Although some standard ML methods are implemented in the **PHASE** package, the emphasis here is on Bayesian techniques as in **BAMBE** (Simon and Larget, 2001) or the widely used **MrBayes** (Ronquist and Huelsenbeck, 2003).

One can observe strong compositional differences when comparing the G+C content between the rRNA genes of different species. For reasons that are detailed in chapter 4, RNA molecules of thermophilic species are G+C rich compared to mesophilic species. This is not compatible with the fourth and fifth assumptions and it has been noticed that this compositional bias can mislead traditional phylogenetic reconstruction techniques. Species can appear grouped according to their nucleotide composition rather than their evolutionary relationships. In chapter 4, new methods and models are proposed to relax these assumptions in the Bayesian framework. The variability of the substitution process over time is accounted for using different substitution models on different branches of the tree. Reversible jump MCMC techniques are used to determine the number of models (and consequently the amount of heterogeneity) needed to fit the data properly.

Based on the analysis of real mitochondrial RNA sequences, it is suggested in chapter 5 that the evolutionary process can be different at sites under strong and weak selection pressure. This is in violation with the third assumption. A substitution model that can account for the fact that the observed nucleotide composition is not the same at slow and fast evolving sites is introduced. This substitution model allows for the variation of the evolutionary process across sites.

1.4 Thesis structure

The structure of the thesis is as follows:

- In **chapter one** the concept of phylogenetic trees that show evolutionary relationships between organisms was introduced. The RNA sequence data used in this thesis were also presented. It is proposed to develop complex evolutionary models to improve our understanding of RNA evolution and to produce more accurate estimates of phylogeny.

- In **chapter two** the main components of current evolutionary models are introduced. Probabilistic substitution models used to describe the evolution of molecular sequences along the branches of the phylogenetic tree are described first. The likelihood function, which is the basis of the inference technique, and its calculation are then explained.

- In **chapter three** the Bayesian approach to phylogenetic inference is introduced along with the Markov chain Monte Carlo technique implemented in the **PHASE** software.

- In **chapter four** a time-heterogenous substitution model is proposed to account for base composition variation over evolutionary time. The reversible jump technique used to perform phylogenetic inferences with this model is described and illustrated with synthetic and real datasets.

- In **chapter five** it is shown that the equilibrium base frequencies of the substitution process also vary across sites. It is shown that it can adversely affect the estimation of parameters when standard substitution models are used. A new method is introduced to account for variation of the evolutionary process across sites.

- In **chapter six** this work is concluded and the contributions are summarized. The weaknesses of the proposed methods and evolutionary models are identified. Future work that could be conducted to bring this research further is proposed.

Chapter 2

Substitution models and the likelihood function

The approach followed to reconstruct evolutionary histories requires a stochastic model of the processes that govern sequence evolution. The two main model components, used by virtually all likelihood-based methods, are described here. First the phylogenetic tree, which depicts the evolutionary relationship for the set of studied species, then the probabilistic substitution model, which describes the nucleotide replacement process. Specific base-pair models, which are appropriate to describe the evolutionary process in RNA helices, are also introduced in this chapter. The pruning algorithm that is used to compute the likelihood of the data and is the basis of all model-based phylogenetic methods mentioned in this thesis is then described. This chapter is concluded with the introduction of more complex evolutionary models and methods that relax the assumptions presented in the introduction and are used in the remainder of this thesis.

2.1 Introduction

As mentioned in section 1.2, any evolving character shared among the set of studied species can be used as a basis for phylogenetic inference. As a result of ongoing sequencing projects, macromolecular data are accumulating fast and new techniques are needed to benefit from increasing computational resources and to mine this useful information. In this thesis, the focus is on molecular data and more specifically on RNA genes. The inferences presented here are consequently based on evolving DNA and RNA sequences but it should be noted that the methods introduced in this chapter, and in this thesis in general, may be extended and applied to amino-acid sequences and discrete morphological characters (Lewis, 2001).

DNA and RNA molecules are nucleotide chains. The replacement of one nucleotide by another (or *mutation*) in the genetic material of an individual is a common event and different individuals within a population usually have differing genetic information (which is known as polymorphism). This appears to be a potential issue when attempting to build species trees from individual sequences. In theory, phylogenetic inference should be carried out using consensus sequences. Consensus sequences are chimeric sequences created from the entire population using the nucleotide carried by the majority of individuals at each position. In other words, they are "average" sequences. Hopefully, variation within a species is limited and can safely be discarded. In practice the sequence of one individual is considered to be representative for the sequence of the population.

Species phylogenies are not directly inferred from mutations in individual sequences but from nucleotide replacements in the global consensus sequence. Indeed, mutations arise regularly and spontaneously in individuals but they do not necessarily persist for more than a few generations. Depending on the relative fitness of the mutant and on chance, a mutation can drift to high frequency and fixation in a population. The replacement of the most common nucleotide in the population by another one is called a nucleotide *substitution*. The stochastic models used to describe this replacement process are naturally called substitution models.

Over time, a consensus sequence may also lose some of its elements (deletion) or have elements inserted (insertion). Although these *indels* also contain a strong phylogenetic signal (see, *e.g.*, Baldauf and Palmer, 1993), the methods developed in this thesis cannot handle them and, consequently, sequence data need to be aligned before analysis (see Figure 2.1). Nucleotides at each column position have to be *homologous* and gaps are inserted to replace the missing nucleotides in some sequences. Extra nucleotides can also be removed for the purpose of phylogenetic inference.

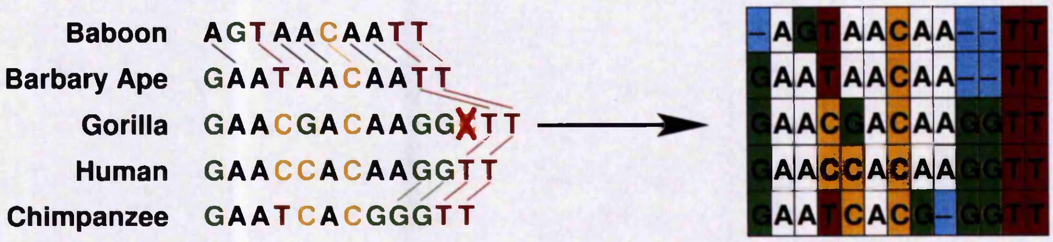


Figure 2.1: Nucleotide alignment procedure.

Alignment is a difficult and important step preceding a phylogenetic analysis. A review of automatic alignment techniques is beyond the scope of this thesis (see, *e.g.*, Wallace et al., 2005; Gardner et al., 2005). For the purpose of this thesis, alignments were refined by eye with reference to the RNA secondary structure and are assumed to be accurate.

2.2 Phylogenetic inference from aligned sequences

2.2.1 Phylogenies

The phylogenetic tree represents the hierarchical relationships arising through evolution among a set of selected species. The species, also called *taxa*, are at the leaves of the tree and have all evolved from a unique common ancestor which is at the root. Leaves are connected to the root by a set of internal nodes linked with branches of different lengths. The internal nodes are bifurcation points between genetically isolated groups or *monophyletic clades* (see Figure 2.2).

Until recently, almost all phylogenetic methods accepted the appropriateness of a tree-like model to describe the evolutionary process but this may be an unwarranted assumption in some cases due to horizontal gene transfer events (HGT) or, perhaps, genome fusions (Rivera and Lake, 2004; Creevey et al., 2004; Kunin et al., 2005). Although one ought to remain cautious, it is relatively safe to assume that nuclear rRNA is free of horizontal transfer because it cannot be functional if it is not transferred with the various ribosomal proteins tightly coupled with it (Woese, 2000). There is growing evidence for HGT of mitochondrial genes in plants (Bergthorsson et al., 2003), but mammalian mitochondrial genes have been extensively studied and HGT is unlikely for most datasets studied in thesis.

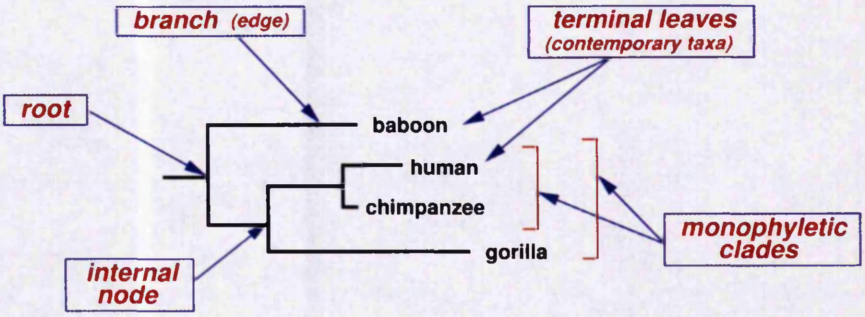


Figure 2.2: A phylogenetic tree: the information conveyed by the tree can be understood by locating the monophyletic clades.

Biologists can choose among a wide range of methods to reconstruct phylogenies and selecting an appropriate method is actually a complex issue. The recent burst of techniques came with an extensive literature but the occasional phylogeneticist will find it rather difficult to penetrate. Three major inference techniques dominate the field: parsimony, distance methods and likelihood methods. Likelihood methods, succinctly introduced in the first chapter, are central to this thesis and are discussed at length later on. Meanwhile, parsimony and distance methods are briefly introduced.

2.2.2 Parsimony methods

Maximum Parsimony methods are based on a different optimality criterion than Maximum Likelihood (ML) methods. Rather than trying to find the tree that yields the highest likelihood, parsimony methods favour the trees that require the fewest number of character changes and rely on a principle widely used in science: simpler explanations are usually preferable to more complex ones. Although there have been many variants (reviewed by Felsenstein, 2004, chap. 7), the simplest parsimony method assumes that all nucleotide changes have equal evolutionary cost and trees are consequently scored according to the minimum number of substitutions necessary to produce the observed sequences. The optimal tree is defined as the tree with the lowest score.

Maximum parsimony has a long history in phylogenetic inference (see, *e.g.*, Camin and Sokal, 1965; Eck and Dayhoff, 1966) and has been, by far, the most widely used method in molecular phylogenetics since Fitch (1971) described an algorithm to compute the minimum number of changes per site for a given tree. **PAUP*** (Swofford, 2003) is an important software that incorporates parsimony methods and which produced a significant share of all the published phylogenies.

Although not clearly stated, parsimony methods are implicitly assuming that changes are rare. This may be a justified assumption over short evolutionary time-scales when the branches of the tree are short, but multiple substitutions will occur at some sites as molecular sequences diverge. Since parsimony cannot correct for superimposed changes, it usually underestimates evolutionary distances. Long-branch attraction is a specific consequence of this issue that was noticed and explained early on (Felsenstein, 1978). Long-branch attraction is a reconstruction artifact where long branches cluster together and share common ancestors regardless of the true underlying evolutionary relationships. This problem cannot be alleviated by increasing the amount of characters studied because the artifact gets more pronounced as the sequence lengths increase.

Parsimony methods are not mentioned in the remainder of this thesis but they are introduced here because they are widely used and are, arguably, the most natural attempt to solve the phylogenetic problem. Even though parsimony can easily be described, the algorithms used to compute the minimum number of changes on a tree and to search for the most parsimonious tree are not so straightforward. In spite of their apparent simplicity, parsimony methods are still computationally expensive. As with the ML optimality criterion, vast numbers of candidate phylogenies have to be evaluated to select the best one(s).

2.2.3 Distance methods

Introduced shortly after parsimony, distance methods are based on a matrix of pairwise evolutionary distances (see, *e.g.*, Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967). In the first step, available data are converted into a symmetric square matrix D of dimension N (N being the number of species) where D_{ij} is the evolutionary distance between the species i and j. The pairwise distances are then used to reconstruct the evolutionary tree. Any reasonable transformation can be used to convert the initial data into a distance measure. Nevertheless, distance methods are known to be better behaved when the distance used is additive, *i.e.*, when the distance between two species is equal to the total sum of the branch lengths that separate them. Markov models of nucleotide substitutions presented later in this chapter can be used to compute a ML estimate of the distance between two species and this functionality has actually been added to **PHASE** in order to be used with any of the RNA substitution models implemented during this project. Since they are relevant to the subject of this thesis, the paralinear/LogDet distances (Lake, 1994; Lockhart et al., 1994) and the distance proposed by Galtier and Gouy (1995) are worth mentionning and will be discussed in chapter 4 because they can successfully account for inequalities of base composition across species.

In the second step, distance measures are used to construct a tree. Many methods are available to construct phylogenies from pairwise distances. The simplest, and fastest, methods are the algorithmic methods that follow a clustering scheme. In these algorithms, closely related taxa are iteratively selected, linked and replaced by their common ancestor until the tree is complete. These clustering algorithms, *e.g.*, Unweighted Pair Group Method with Arithmetic mean (UPGMA, Sokal and Michener, 1958) or Neighbour Joining (NJ, Saitou and Nei, 1987), have a long history and are still widely used today, due to their ability to deal with very large datasets. They have recently been improved by accounting for the noise in the evolutionary distances (Gascuel, 1997; Bruno et al., 2000).

Distance reconstruction methods described above apply their algorithm to produce a tree from a distance matrix and the phylogeny is completely defined by the algorithm. Other distance methods exist which are more firmly grounded in a statistical framework. These alternatives make use of an optimality criterion, which implies that many topologies have to be evaluated in order to find the best one. The least squares method and its weighted and generalized variants attempt to minimize the differences between the observed measures in the pairwise distance matrix and the corresponding branches of a candidate tree (Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967). The phylogeny, with associated branch lengths, that fits the matrix with the least residual error is considered the best. The minimum evolution criterion (Rzhetsky and Nei, 1992) is another related, but slightly different, optimality criterion. Once branches have been fitted with least squares on candidate phylogenies, minimum evolution chooses the shortest phylogeny (*e.g.*, with the smallest sum of branch length) rather than the phylogeny that best fits the distance data.

Obviously, there is an inevitable and massive loss of information when the original dataset, molecular sequences in our case, is squashed into a $N \times N$ matrix. However, contrarily to intuition, experience and computer simulations have shown that pairwise distance methods preserve most of the information and often behave remarkably well in practice. Nevertheless, these methods do not provide great insight into the process of evolution.

2.3 Nucleotide substitution models

2.3.1 Markov models of nucleotide substitution

DNA substitution models are designed to model nucleotide substitution in homologous DNA strands. Replacements within these sequences are described and modelled by a 4-state Markov process, each state represents one of the base found in DNA molecules: **A**denine, **G**uanine, **C**ytosine and **T**hymine (see Figure 2.3).

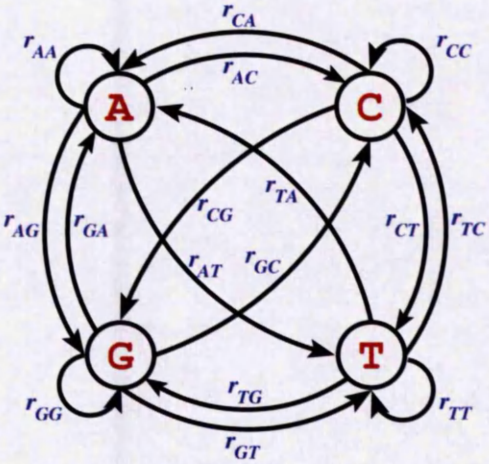


Figure 2.3: A Markov model of nucleotide substitution.

Phylogenetic inference is usually done from the mRNA complementary transcripts rather than the original genes. In RNA molecules, **T**hymine is replaced by **U**racil. This is of little consequence for our substitution models and it is assumed that **T** and **U** are equivalent and interchangeable from now on.

This 4-state Markov process is completely specified by its rate matrix, which contains the rates of substitution between the four bases. Following the notation of Swofford et al. (1996),

$$Q = \begin{array}{c c} & \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{array}{c} A \\ C \\ G \\ T \end{array} & \begin{pmatrix} r_{AA} & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & r_{CC} & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & r_{GG} & r_{GT} \\ r_{TA} & r_{TC} & r_{TG} & r_{TT} \end{pmatrix} \end{array} \quad . \tag{2.1}$$

Each element r_{ij} represents the instantaneous rate of substitution from nucleotide i to nucleotide j if $i \neq j$. $r_{ij} \times dt$ is the probability of change from i to j in an infinitesimal amount of time dt. The probability of i **not** changing into another nucleotide in time dt is consequently $1 - \sum_{j \neq i} r_{ij} dt$ but the diagonal of Q is defined by

$$\forall i, \qquad r_{ii} = -\sum_{j \neq i} r_{ij} \quad , \tag{2.2}$$

for convenience so that

$$dP(t) = P(t) \times Q dt \quad , \tag{2.3}$$

where $P(t)$ is the matrix of substitution probability, which gives, for each couple of nucleotides i,j the probability that state i is replaced by state j in time t (or probability that state i is still state i in the case $i = j$). Solving differential equation (2.3) with $P(0) = I$ gives:

$$P(t) = e^{Qt} \quad . \tag{2.4}$$

The exponentiation in (2.4) is done by diagonalising Q with its eigenvectors and eigenvalues. **PHASE** uses the **BLAS/LAPACK** FORTRAN libraries to perform this computation.

As mentioned in the introduction, it is traditionally assumed that the substitution process is stationary or at equilibrium. In such a case, a vector of equilibrium frequency parameters, constant over time, can be defined. These frequency parameters ($\mathbf{\Pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$) add up to one. Notice that for any column vector X representing an initial marginal distribution over the nucleotides, $\lim_{t \to +\infty} X^T \times P(t) = \mathbf{\Pi}$ (see Figure 2.4). After an infinite amount of time, the composition, averaged over the complete sequence, converges towards the equilibrium frequencies. Based on the stationary assumption, it is also supposed that the process was already at equilibrium at $t = 0$ (the root of the tree).

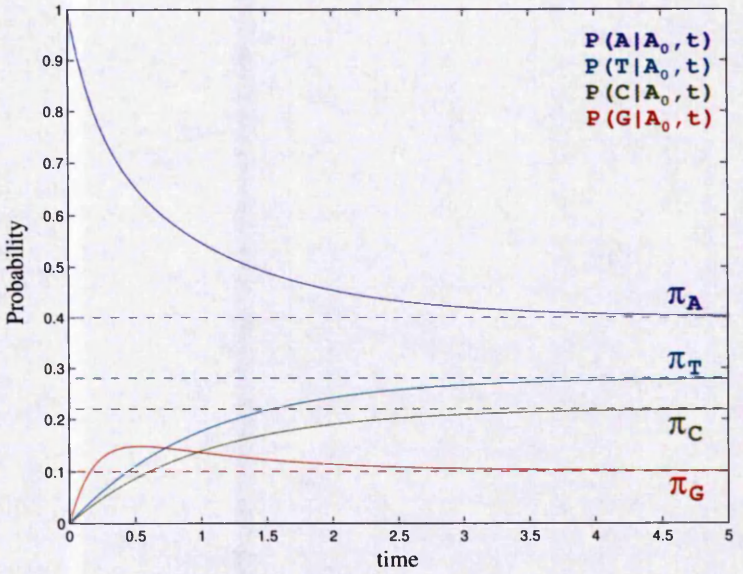


Figure 2.4: Evolution of a nucleotide over time: this graph returns the probability of a nucleotide being one of the four bases at different time-points, given that the initial nucleotide at $t = 0$ was an **A**. Probabilities converge to the equilibrium frequencies $\{\pi_A, \pi_C, \pi_G, \pi_T\}$. The substitution model used is the HKY model (described in 2.3.2) with the transition/transversion ratio set to 10.

Most substitution models also assume that, averaged over the whole sequence, the flux from base i to base j is equal to the flux from j to i. This assumption is enforced using the detailed balance equation:

$$\forall (i,j)/i \neq j \qquad \pi_i r_{ij} = \pi_j r_{ji} \quad . \tag{2.5}$$

The substitution models considered in this thesis belong to this class of models and are said to be *time-reversible.* Reversibility is first and foremost a mathematical convenience but it is probably not far from the reality. One consequence of time-reversibility is that the substitution model cannot be used to determine the direction of evolution. With two sequences at opposite ends of a branch, it cannot be decided which is the ancestral sequence and which is the derived sequence. For our problem, the important consequence of time-reversibility is that substitution models can be used to infer phylogenies but cannot be used to decide where to root them. Extra knowledge is needed to position the root when such substitution models are used.

By introducing a set of parameter $R = \{\rho_{ij}\}$ and defining:

$$\forall(i,j)/i \neq j, \qquad r_{ij} = \rho_{ij}\pi_j \quad , \tag{2.6}$$

then the detailed balance equation (2.5) can automatically be satisfied for a symmetric choice of R. The elements of R are called exchangeability parameters in this thesis.

2.3.2 Standard nucleotide models

Building on the previous equations, the most general time-reversible nucleotide substitution model, known as GTR or REV, can be expressed by the following rate matrix:

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & \rho_{AC}\pi_C & \rho_{AG}\pi_G & \rho_{AT}\pi_T \\ \rho_{AC}\pi_A & - & \rho_{CG}\pi_G & \rho_{CT}\pi_T \\ \rho_{AG}\pi_A & \rho_{CG}\pi_C & - & \rho_{GT}\pi_T \\ \rho_{AT}\pi_A & \rho_{CT}\pi_C & \rho_{GT}\pi_G & - \end{pmatrix} \end{array} . \tag{2.7}$$

This model was introduced by Lanave et al. (1984) but was not extensively used until recently. Other nucleotide models commonly used are biologically motivated simplifications of this general model. Even though these alternative models are presented here as constrained versions of the GTR model, history proceeded in the opposite direction and models became increasingly complex as assumptions were relaxed and computational resources improved.

JC69 is the simplest 4-state model of DNA sequence evolution which was proposed by Jukes and Cantor (1969). In this model, all nucleotides are assumed to be interchangeable and the transitions of the Markov model are all equal. This corresponds to all ρ_{ij} being equal in (2.7) and $\pi_A = \pi_C = \pi_G = \pi_T = 25\%$.

From a biochemical point of view, the four nucleotides present some obvious differences and one can distinguish the purines (**A** and **G**) and the pyrimidines (**C** and **T**/**U**) by their number of heterocyclic compounds. The K80 model (Kimura, 1980) accounts for these two families and improves on the JC69 model by using two different substitution rates. The *transition* rate (in the biological sense, purine→purine and pyrimidine→pyrimidine) is no longer assumed to be the same as the *transversion* rate (purine→pyrimidine and pyrimidine→purine) and two independent exchangeability parameters are used in (2.7) ($\rho_{transition} = \rho_{AG} = \rho_{CT}$ and $\rho_{transversion} = \rho_{AC} = \rho_{AT} = \rho_{CG} = \rho_{GT}$, see Figure 2.5). In the K80 model, nucleotide frequencies are still assumed to be equal to 25% and this assumption was relaxed by Hasegawa et al. (1985) in the HKY85 model.

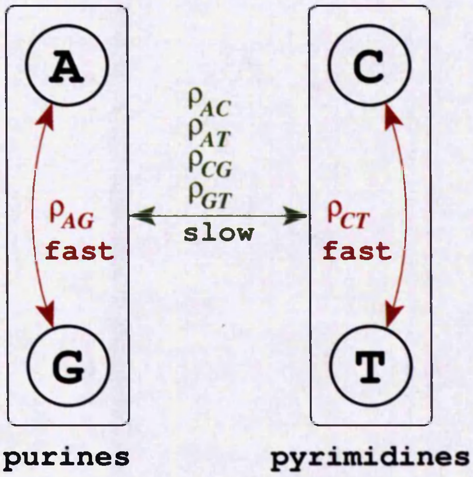


Figure 2.5: The K80 and HKY85 substitution models: all transversion rates are equal $\rho_{AC} = \rho_{AT} = \rho_{CG} = \rho_{GT} = \rho_{transversion}$ and all transition rates are equal $\rho_{AG} = \rho_{CT} = \rho_{transition}$. The **A** and **G** nucleotides are chemically similar and fall into the group known as the purines while **C** and **T** are known as the pyrimidines. Nucleotide substitutions within groups (transitions) are much more frequent than substitutions between groups (transversions) and typically $\rho_{transition}/\rho_{transversion} > 1$. Unlike the HKY85 model, the K80 model assumes that nucleotide frequencies are equal.

Clearly, one could still come up with a different set constraints on R and Π and propose a "new" substitution model. Few of the 203 possible time-reversible variations of R have been named so far (Huelsenbeck et al., 2004). Since biological arguments have already been exploited, the real challenge lies now on the statistical justification of any particular choice. Consequently, only one last substitution model is introduced here because it is used later on. This model was proposed by Tamura and Nei (1993) and is a special case of the GTR model (see matrix (2.7)) where all transversion rates are equal:

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & \rho_{trans}\pi_C & \rho_{AG}\pi_G & \rho_{trans}\pi_T \\ \rho_{trans}\pi_A & - & \rho_{trans}\pi_G & \rho_{CT}\pi_T \\ \rho_{AG}\pi_A & \rho_{trans}\pi_C & - & \rho_{trans}\pi_T \\ \rho_{trans}\pi_A & \rho_{CT}\pi_C & \rho_{trans}\pi_G & - \end{pmatrix} \end{array} . \qquad (2.8)$$

In other words, the TN93 model can also be defined as a generalisation of the HKY85 model, where the two transition rates ρ_{AG} and ρ_{CT} are not constrained to be equal.

2.4 RNA genes and base-pair substitution models

2.4.1 RNA molecules and compensatory substitutions

In 1953, Watson and Crick described the structure of the DNA molecule and suggested that it is composed of two intertwined nucleotide chains. The two strands are held together by hydrogen bonds between the Watson-Crick pairs **A-T** and **C-G** and their nucleotide sequences are perfectly complementary.

Unlike DNA molecules, RNA molecules are single stranded and they naturally fold into a compact and complex structure when left in their environment. Although numerous interactions are known to play a critical role in the final three-dimensional shape or *tertiary structure* of these ribozymes, the configuration of the molecule is mainly determined by the Watson-Crick bonding interactions between distant nucleotides (*secondary structure*). In RNA molecules, helices are formed intramolecularly and the folding process produces regions of paired nucleotides or *stems*, interspersed with unpaired strands loosely called *loops* in this thesis (see Figure 2.6 and Figure 2.7).

In RNA stems, the helical structure is maintained by traditional Watson-Crick pairs but one can also notice non-canonical base-pairs. Purine-purine and pyrimidine-pyrimidine hydrogen bondings are unlikely but the two other purine-pyrimidine interactions (predominantly **G-U** pairs and, to a lesser degree, **A-C**

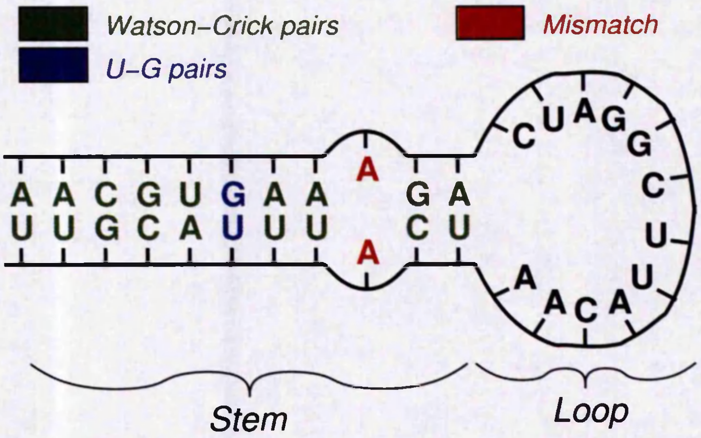


Figure 2.6: Schematic representation of the RNA secondary structure.

pairs) can, in some cases, be almost as stable as the standard Watson-Crick pairs.

Ribosomal RNAs and transfer RNAs have an important role and their structure cannot easily be disrupted without impact on their function and potentially lethal consequences. Therefore, selection pressure is acting to maintain the secondary structure and the stems. Yet, the *primary structure* (the unidimensional nucleotide sequence) can still vary and, in fact, we observe that RNA helical regions are quite variable in sequence. The bases used are usually not as important and substitutions are possible as long as the secondary structure is preserved.

The secondary structure is unchanged when complementary substitutions occur in the DNA gene coding for the RNA molecule. From an individual sequence viewpoint, this mechanism is a two-step process. A mutation on one side of a pair (say **G-C** to **G-U**) slightly disturbs the helicoidal structure because the **G-U** bond is the least stable thermodynamically. However, a second mutation on the opposite site (**G-U** to **A-U**) can fully restore the pairing ability. Recall that phylogenetic inference is based on substitutions and not on individual mutations. From the population viewpoint, evolution in stems can occur either by two successive substitutions or by a simultaneous compensatory substitution (Stephan, 1996; Higgs, 1998). In the first case, the slightly deleterious **G-U** pair drifts to fixation in the population and is replaced by the **A-U** pair later on. In the second case the selection against intermediate mutants is too strong and the **G-U** pair is kept at low frequency in the population until a second mutation takes place in one of the sporadic mutant (see Figure 2.8).

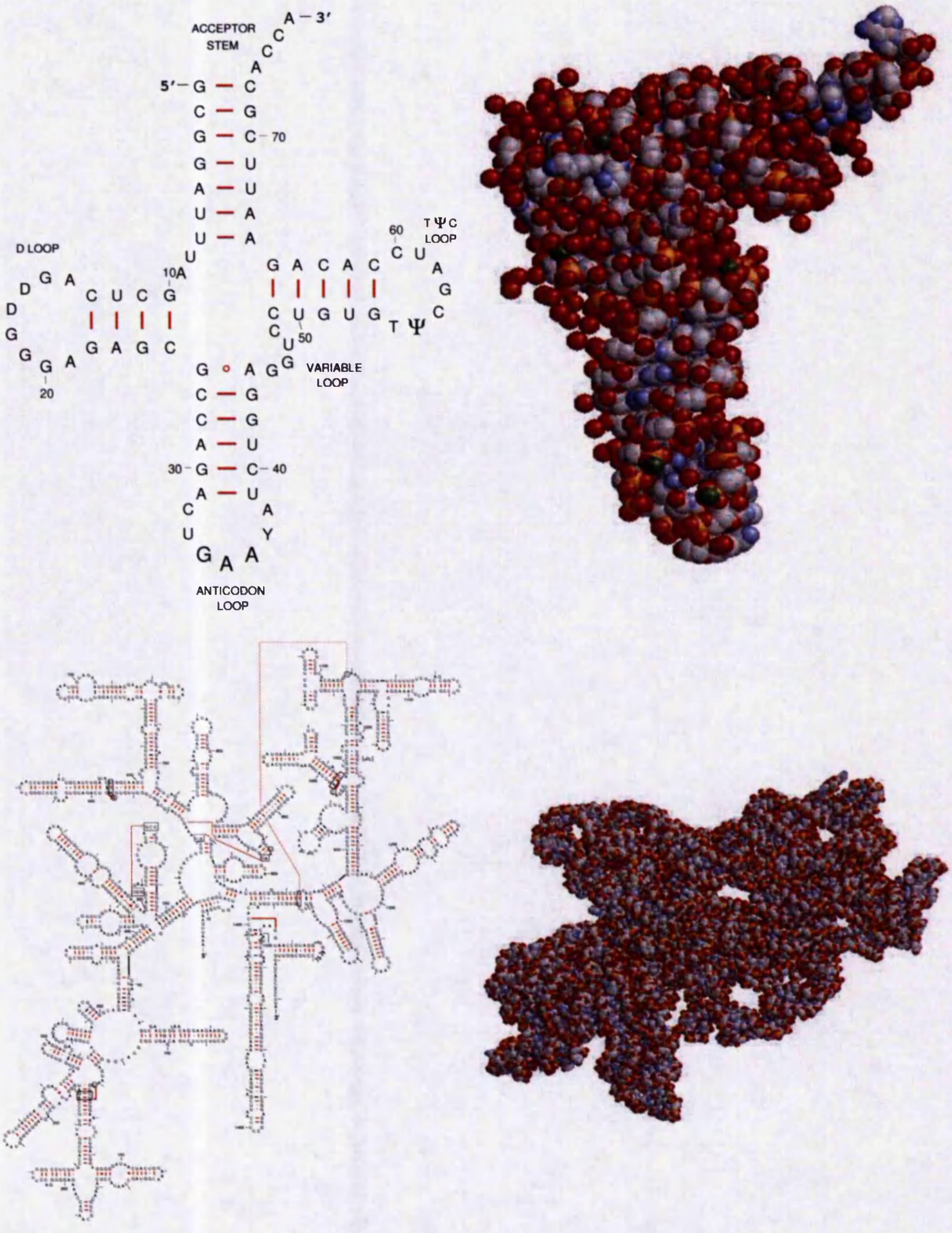


Figure 2.7: The secondary and tertiary structure of common RNAs; top: tRNA-Phe (Yeast), bottom: rRNA Small SubUnit (E.Coli); secondary structures on the left are from the Comparative RNA Web Site (http://www.rna.icmb.utexas.edu/), tertiary structure on the right were realised with rasmol, pdb structure files 1EVV and 1PNS.

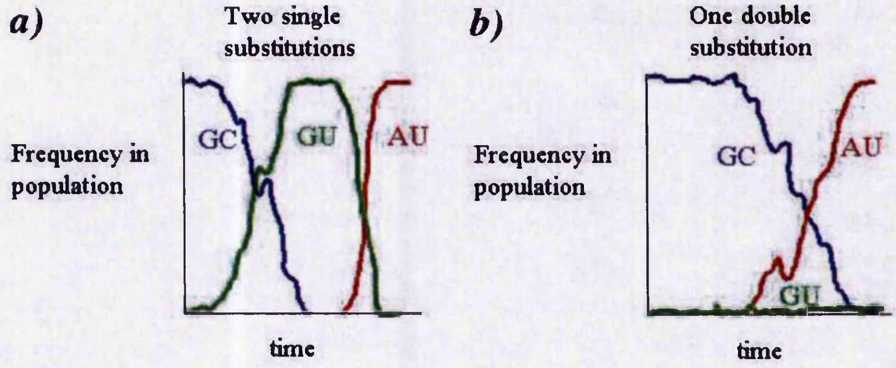


Figure 2.8: The compensatory substitution mechanism: a) the intermediate **G-U** pair reaches fixation before a second substitution occurs, b) the deleterious **G-U** pair is kept at a low frequency until a second mutation restores the pairing ability and the new **A-U** pair gets fixed. Figure courtesy of Paul Higgs.

2.4.2 RNA substitution models

To model the evolution of nucleotides in stems, and to use RNA genes in phylogenetic inference, one could simply use a standard DNA substitution model as described previously. However, non-canonical pairs are not common (approximately 10% to 15% of the base-paired sites), and one side of the double helix can almost be deduced from the other side. Using a DNA model with RNA stems is consequently almost equivalent to accounting for the same data twice and usually leads to overconfidence in the results. In the worst case, one can find strong support for an incorrect phylogeny (Tillier and Collins, 1995; Jow et al., 2002; Galtier, 2004). In practice, most RNA phylogenies are done with standard four-state DNA substitution models even though the violation of the assumption of independence among sites is known to bias the results. In this thesis, substitution models that properly account for the secondary structure of RNA molecules are favoured.

Since the evolutionary pressure is acting on the structure rather than on the nucleotides themselves, RNA stems are quite variable in their sequences but the stems are relatively well conserved over evolutionary time. In an alignment of homologous RNA genes, most columns can consequently be unambiguously characterised as either being independent of being part of a pair. The assumption of independence can then partially be alleviated by simultaneously considering the two co-evolving nucleotide positions constituting a pair. Independent positions in RNA loops can be handled with one of the standard 4-state models described in section 2.3 and new probabilistic substitution models have to be proposed for the stems.

A variety of substitution models have been proposed to model RNA sequence evolution (Savill et al., 2001). These base-pair models differ from the previous nucleotide models in that the pair is now considered as the elementary unit of evolution. Since there are sixteen possible base-pairs (**A-U** and **U-A** are considered as different states), the most general base-pair substitution models presented here naturally have sixteen states even though the six stable base-pairs (four Watson-Crick pairs plus the **U-G/G-U** wobble base-pairs) constitute the majority of sites in RNA helices. In a typical alignment, 5% of the pairs belong to the class grouping the ten other possible combinations (mismatch pairs). Mismatch pairs can be caused by unconventional pairing interactions but they are generally related to modifications of the secondary structure over evolutionary time, which is not accounted for by current models.

A general sixteen state time-reversible model of base-pair substitutions can easily be constructed by analogy to the most general DNA substitution model presented above. However, this parameter-rich model, with sixteen frequency parameters and 120 exchangeability parameters, would be of little use in practice. This is not as much a computational issue as a data issue. Indeed, a reasonably sized RNA dataset hardly provides enough information to estimate all these parameters properly, especially those associated with rare pairs and rare substitution events. Since data concerning mismatches is scarce, one straightforward solution is simply to remove them from the alignment and use a six-state model (WC pairs + UG/GU) with the paired sites that are conserved across species and over evolutionary time (Tillier, 1994). To avoid wasting information, one can also consider the two columns of the unstable pair as independent, and treat them as such, rather than removing them.

When large alignments (in terms of number of species) are studied, the probability of a pair being conserved in all species is getting lower. Another solution is consequently not to disregard mismatches completely but to lump them into a single state (Tillier and Collins, 1998). The original dataset is recoded and the ten mismatch pairs are lumped into a seventh state **M-M**. Note that a similar technique (**RY**-coding) can be used with standard DNA models. Reducing the four nucleotides (**A**, **C**, **G** and **T**) to purines (**R**) and pyrimidines (**Y**) was found to reduce the biasing effects of saturation and compositional differences with standard phylogenetic methods (Harrison et al., 2004).

RNA models implemented in **PHASE** have consequently six, seven or sixteen states. Within their family, these different sets of models have been compared against real data by Savill et al. (2001) using likelihood-ratio tests, Akaïke's Information criterion (1974) and Cox's statistical test (1962). See Posada and Crandall (2001) and Goldman (1993) for the application of these tests to the phylogenetic problem. At this point, it might be better to classify these RNA models into two classes, which are not related to their number of states.

The first class contains RNA substitution models that are heavily inspired by their DNA counterparts. In these substitution models, transitions (in the probabilistic sense) that require the simultaneous change of the two nucleotides of a pair are not allowed. Indeed, according to the DNA Markov models presented previously, the probability of two nucleotides simultaneously changing in an infinitesimal amount of time dt is a negligible term in dt^2. The models proposed by Schöniger and von Haeseler (1994), and Rzhetsky (1995) belong to this class where exchangeability parameters for double substitutions are null. These two models account for the difference in the selective fitness of the sixteen pairs by using different equilibrium frequencies and the six common pairs naturally have higher equilibrium frequencies. Muse (1995) used a different parameterisation for his RNA substitution model and incorporated a selective term λ for the preferential attachment to the Watson-Crick base-pairs. With $\lambda = 1$, this model reduces to the standard HKY85 4-state model.

The second class of RNA models, introduced by Tillier (1994), does not assume that the instantaneous rate of double substitutions is null. While this would make little sense for a mutation model used at the individual sequence level, this can easily be justified at a population level, as pointed out earlier (see Figure 2.8). More importantly, models that allow for double substitutions have been shown to be superior when applied to the evolution of real RNA sequences (Tillier and Collins, 1998; Savill et al., 2001).

Many RNA models have been implemented recently in **PHASE** and are described in the manual of the software. For the purpose of this thesis, introducing the 7A and the 7D models suffice. As indicated by their name, these two models are seven-state models (with the ten mismatches lumped into a single MM state).

The 7A model is the most general time-reversible model with seven states and can be constructed analogously to the 4-state GTR model of nucleotide evolution introduced in section 2.3. It has seven frequency parameters and 21 exchangeability parameters (and consequently allows for double substitutions). The 7D model, also known as the OTRNA model (Tillier and Collins, 1998), is a biologically motivated simplification of the 7A model where some exchangeability parameters that were independent in 7A are constrained to be equal (see Figure 2.9). The 7D model is completely specified by the following rate matrix:

$$Q = \begin{array}{c} \\ AU \\ GU \\ GC \\ UA \\ UG \\ CG \\ MM \end{array} \begin{array}{c} \begin{array}{ccccccc} AU & GU & GC & UA & UG & CG & MM \end{array} \\ \begin{pmatrix} * & \alpha_s\pi_{GU} & \alpha_d\pi_{GC} & \beta\pi_{UA} & \beta\pi_{UG} & \beta\pi_{CG} & \gamma\pi_{MM} \\ \alpha_s\pi_{AU} & * & \alpha_s\pi_{GC} & \beta\pi_{UA} & \beta\pi_{UG} & \beta\pi_{CG} & \gamma\pi_{MM} \\ \alpha_d\pi_{AU} & \alpha_s\pi_{GU} & * & \beta\pi_{UA} & \beta\pi_{UG} & \beta\pi_{CG} & \gamma\pi_{MM} \\ \beta\pi_{AU} & \beta\pi_{GU} & \beta\pi_{GC} & * & \alpha_s\pi_{UG} & \alpha_d\pi_{CG} & \gamma\pi_{MM} \\ \beta\pi_{AU} & \beta\pi_{GU} & \beta\pi_{GC} & \alpha_s\pi_{UA} & * & \alpha_s\pi_{CG} & \gamma\pi_{MM} \\ \beta\pi_{AU} & \beta\pi_{GU} & \beta\pi_{GC} & \alpha_d\pi_{UA} & \alpha_s\pi_{UG} & * & \gamma\pi_{MM} \\ \gamma\pi_{AU} & \gamma\pi_{GU} & \gamma\pi_{GC} & \gamma\pi_{UA} & \gamma\pi_{UG} & \gamma\pi_{CG} & * \end{pmatrix} \end{array} .$$

2.5 Phylogenetic trees and branch lengths in the likelihood framework

Throughout this thesis *rooted* and *unrooted* trees are mentioned. It turns out that the analytical techniques used for tree reconstruction do not always allow for an unambiguous placement of the root. This is related to the fact that some evolutionary models are time-reversible as already mentioned. It is therefore not always possible to infer the position of the earliest point in time, the common ancestor of all species, in a phylogenetic network (see Figure 2.10).

Typically, the most important result of a phylogenetic analysis is the pattern of branching, or tree topology. Nevertheless, when pairwise evolutionary distances have been described at the beginning of this chapter, the notion of branch lengths has also been introduced. A link is now established between the lengths of the branches and the substitution models described in sections 2.3 and 2.4.

Intuitively, branch lengths should be a measure of time but they are actually a measure of evolutionary distance. The elongation between two nodes is the

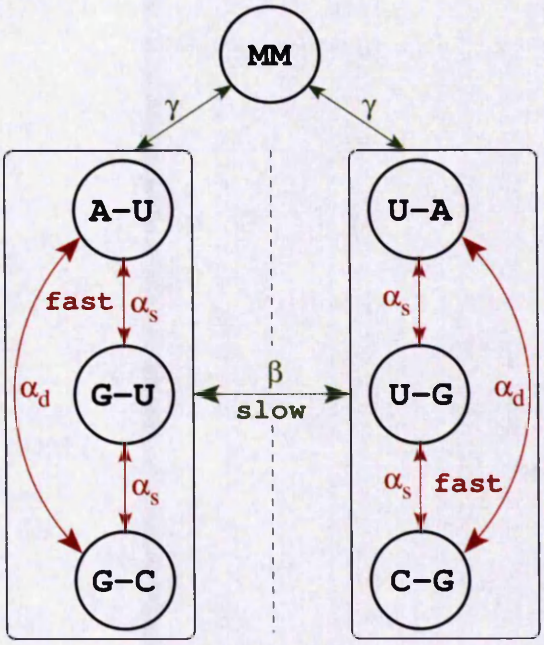


Figure 2.9: The 7D (OTRNA) substitution model (Tillier and Collins, 1998): rapid interchange occurs within the two groups of states (single transitions and double transitions), whereas interchange between the two groups is rare (double transversions).

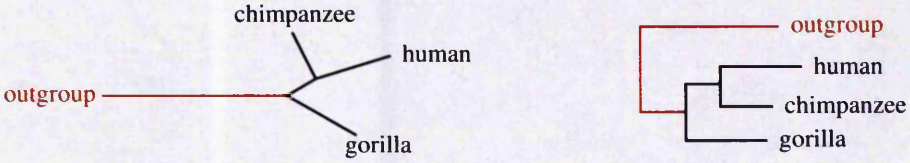


Figure 2.10: Two equivalent representations of the same unrooted trees. When the phylogenetic method used cannot locate the root, an outgroup species, known to be genetically isolated from the others, is traditionally used to restore the tree-like shape.

expected number of substitutions per site along the branch. A large branch indicates that a large number of substitutions separate the two sequences at its incident nodes. The evolutionary distance is related to physical time by a simple relation:

$$dL = r(t)dt \quad , \tag{2.9}$$

where $r(t)$ is the rate of evolution at time t. Since $r(t)$ is unknown, it is not possible to deduce time from branch lengths without further assumptions. Most phylogenetic methods do not tease apart the evolutionary rate and the time since one can only infer their product, which is an evolutionary distance, from the available data.

Recall that in equation (2.4), the probability of a change $P(t)$ was given as a function of $Q \times t$. The evolutionary distance between two sequences can be large because they are separated by a long period of time or, indistinguishably, because the rates of substitution (*i.e.*, the terms in the matrix Q) are high. To resolve the identifiability issue, the matrix Q is scaled by a factor μ so that the average substitution rate of the substitution model:

$$E = \mu \times \sum_{i=1}^{nb_{states}} \sum_{j \neq i} \pi_i r_{ij} \quad , \tag{2.10}$$

is equal to one. This makes the length of a branch equal to the expected number of substitutions per site along that branch. Since Q is scaled before being used, proportional sets of transition rates $\{r_{ij}\}$ are equivalent. The identifiability issue is now inside the matrix and it is necessary to add an extra constraint on the exchangeability parameters. One of the exchangeability parameters can be fixed to 1 and used as a reference. Alternatively, one can enforce the constraint $\sum_{i=1}^{nb_{exch}} \rho_i = 1$ where nb_{exch} is the number of free exchangeability parameters in the substitution model. Both methods are used in this thesis because this constraint was recently changed in the **PHASE** software. As will be explained in chapter 3, this particular choice of parameterization is of no consequence in the ML framework but can have an effect in Bayesian inference.

In general, no assumption is made about $r(t)$ in equation (2.9) and reference to physical time is dropped out. Nevertheless, it might be reasonable to assume that $r(t)$ is constant over time (global molecular clock) or smoothly varying (relaxed molecular clock). In practice, these two assumptions are used with *ultrametric* trees. In an ultrametric tree, branch lengths are directly proportional to time span and not necessarily to the amount of change. Consequently, the terminal branches leading to contemporary species stop simultaneously at time $t = 0$ (see Figure 2.11). The root is uniquely defined as the point being at equal distance from all leaves and traditional methods can consequently position the root without ambiguity when the (relaxed) molecular clock is assumed.

Some methods implemented in **PHASE** have been adapted to handle ultrametric trees when a global molecular clock is assumed. Since this assumption

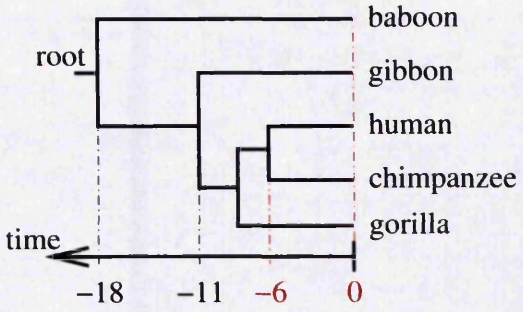


Figure 2.11: An ultrametric tree: contemporary species are on the same timeline $t = 0$. With an appropriate calibration point, for instance the human/chimpanzee split approximately 6 million years ago, it becomes possible to date other speciation events.

is not appropriate for the genes and datasets used in this thesis, these algorithms and molecular clocks in general are not described further in this thesis. However, note that these methods are drawing considerable interest. Time information is crucial to discover the connections between Earth's history and biological evolution, for instance the rise of oxygen levels 2.2 billion years ago and the origin of photosynthetic Cyanobacteria. A reliable dating of some specific speciation events is probably more useful than the overall pattern of branching in some research. The global clock hypothesis is incompatible with most datasets (Tajima, 1993) and local clock methods have been devised in the ML and Bayesian frameworks (Sanderson, 1997; Thorne et al., 1998; Kishino et al., 2001; Huelsenbeck et al., 2000). These methods are popular because they can estimate divergence times without assuming rate constancy. Recovering the evolutionary timescale of life using phylogenetic methods is an active, and controversial, research area (Hedges and Kumar, 2003; Graur and Martin, 2004).

2.6 The likelihood function and the pruning algorithm

2.6.1 The pruning algorithm: simple case

Substitution models that are used to compute probabilities that nucleotides and paired-sites change over time have been defined. It is described here how these substitution models can be used in practice to "score" candidate phylogenies with a likelihood value.

Given a phylogenetic tree topology τ and its associated set of branch lengths ν, one can compute the likelihood of a set of aligned sequences X. Depending on the sequences studied, one (or more) of the substitution models described above is naturally a part of this evolutionary model and the likelihood is also a function of θ, the set of free parameters of this substitution model (*e.g.*, the frequency and exchangeability parameters that are allowed to vary). The likelihood is the probability of the sequence data given the *generative* evolutionary model: $P(X|\tau,\nu,\theta)$. In other words, the likelihood can be understood as the probability that a specific phylogeny and set of substitution parameters have generated the observed sequences. This probability is obviously very low because a given evolutionary model (phylogeny + substitution model) can generate many datasets with almost equal probabilities. Nevertheless, this probability should not be confused with the more intelligible probability that the model is correct given the observed dataset $P(\tau,\nu,\theta|X)$.

Felsenstein (1981) described a practical algorithm to compute the likelihood function. For the sake of clarity, this algorithm is described here in its simplest form with a set of aligned DNA sequences. Since it is assumed that different sites are evolving independently, it turns out that the probability of the data given the evolutionary model can be computed site by site because the overall likelihood is a product of these terms:

$$\begin{aligned} L =& P(X|\tau,\nu,\theta) \\ =& \prod_{j=1}^{nb_{sites}} P(X_j|\tau,\nu,\theta) \quad , \end{aligned} \tag{2.11}$$

where X_j is the data at the jth site. It is recalled that when an RNA gene is used, the two elements of a pair are considered as a single site and the independence assumption is thus not an issue in this specific case. Felsenstein's algorithm will consequently be explained with the computation of the likelihood at a single site.

It is known that interactions between neighbouring sites influence the evolution of molecular sequences and some substitutions are known to be favored because of the context. For instance, there is an excess of **C→T** transitions in **CpG** dinucleotides. It is also known that adjacent pairs in RNA stems are stabilized by *stacking interactions* and influence each other. Thus, independence is a simplifying assumption which is (slightly) violated with real data. Nevertheless, practical computation does not appear feasible without it (but see Siepel and Haussler, 2004; Jojic et al., 2004).

The pruning algorithm is explained with a rooted tree (Figure 2.12) and it is actually necessary to choose a root to apply this recursive algorithm. However, it can be shown that this particular choice does not affect the likelihood value when the model used is time-reversible (pulley principle, Felsenstein, 1981) and, as previously noted, the likelihood function cannot be used to position the root in general.

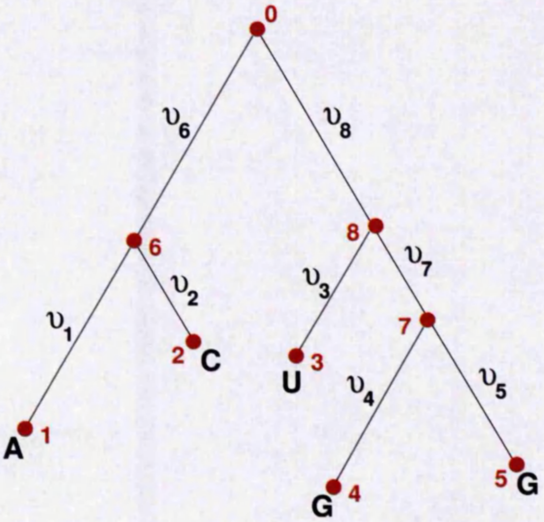


Figure 2.12: The phylogenetic tree that is used in the discussion. Numbers identify internal nodes and leaves, ν_i are the branch lengths.

If we assume that the nucleotides at ancestral nodes $\{s_i\}$ are known, the likelihood at a site can easily be written according to the states observed for a given site at the leaves of the tree. For the tree represented in Figure 2.12, this likelihood would be:

$$L'_j = P(s_0, s_6, s_7, s_8, \mathbf{A}, \mathbf{C}, \mathbf{U}, \mathbf{G}, \mathbf{G} | \tau, \nu, \theta) \quad . \tag{2.12}$$

The Markov substitution models described previously can be used to decompose this expression using the probabilities of change in each tree segment:

$$\begin{aligned} L'_j = \quad & \pi_{s_0} \times P_{s_0 \to s_6}(\nu_6) \times P_{s_0 \to s_8}(\nu_8) \\ & \times P_{s_6 \to \mathbf{A}}(\nu_1) \times P_{s_6 \to \mathbf{C}}(\nu_2) \\ & \times P_{s_8 \to \mathbf{U}}(\nu_3) \times P_{s_8 \to s_7}(\nu_7) \\ & \times P_{s_7 \to \mathbf{G}}(\nu_4) \times P_{s_7 \to \mathbf{G}}(\nu_5) \quad , \end{aligned} \tag{2.13}$$

where π_{s_0} is the prior probability of having nucleotide s_0 at the root and is also the equilibrium frequency of the state s_0 given by the substitution model since the process is assumed to be stationary. $P_{\mathbf{Y} \to \mathbf{Z}}(l)$ is the probability that nucleotide $\mathbf{Y}$ is substituted by nucleotide $\mathbf{Z}$ along a branch of length l.

Of course, ancestral states are not known and they are not a part of the dataset. The likelihood of the data is actually $P(\mathbf{A}, \mathbf{C}, \mathbf{U}, \mathbf{G}, \mathbf{G}|\tau, \nu, \theta)$ which is computed by summing over all the possible assignments for the internal nodes:

$$L_j = P(\mathbf{A}, \mathbf{C}, \mathbf{U}, \mathbf{G}, \mathbf{G}|\tau, \nu, \theta) = \sum_{s_0} \sum_{s_6} \sum_{s_7} \sum_{s_8} L'_j \quad . \tag{2.14}$$

Without further simplifications, the likelihood would not be tractable. Moving the summation signs inwards in equation 2.13 leads to considerable economy in terms of computation:

$$\begin{aligned} L_j = \sum_{s_0} \pi_{s_0} \times \Bigg\{ \sum_{s_6} \Big[P_{s_0 \to s_6}(\nu_6) \times P_{s_6 \to \mathbf{A}}(\nu_1) \times P_{s_6 \to \mathbf{C}}(\nu_2) \Big] \times \sum_{s_8} \Big[P_{s_0 \to s_8}(\nu_8) \\ \times P_{s_8 \to \mathbf{U}}(\nu_3) \times \sum_{s_7} \big(P_{s_8 \to s_7}(\nu_7) \times P_{s_7 \to \mathbf{G}}(\nu_4) \times P_{s_7 \to \mathbf{G}}(\nu_5) \big) \Big] \Bigg\} \quad . \end{aligned} \tag{2.15}$$

As Felsenstein (1981) pointed out, the pattern of parenthesis in equation 2.15 $\Big\{ [\,] \, [(\,)] \Big\}$ bears an interesting relationship to the pattern of branching in the tree $\Big\{ [1, 2], [3, (4, 5)] \Big\}$. It turns out that the likelihood can be computed efficiently by starting at the leaves of the tree and moving towards the arbitrarily chosen root. For each site j and each internal node k, one can define the conditional probability of a subtree $L_{j,k}(s_k)$, which is the likelihood of the data under this internal node k assuming that the ancestral nucleotide at this node is s_k.

The conditional likelihood at a node is defined recursively from the conditional likelihoods of its descendants (see 2.15). If l and m are the immediate descendants of k then:

$$L_{j,k}(s_k) = \sum_{s_l} P_{s_k \to s_l}(\nu_l) L_{j,l}(s_l) \times \sum_{s_m} P_{s_k \to s_m}(\nu_m) L_{j,m}(s_m) \quad . \tag{2.16}$$

The overall likelihood at a site is defined from the conditional probability at the root:

$$L_j = \sum_{s_0} \pi_{s_0} L_{j,root}(s_0) \quad . \tag{2.17}$$

The conditional likelihoods at the tips, which initialise the recursion, are naturally:

$$L_{j,t}(s_t) = \begin{cases} 1 & \text{if } s_t \text{ is the observed nucleotide for the species,} \\ 0 & \text{otherwise.} \end{cases}$$

One does not always observe a nucleotide at a tip. Recall from Figure 2.1 that some gaps have been inserted to produce the final alignment. Also, sequencing techniques are not perfect and some ambiguities can remain in the final alignment. For instance, if a purine **R** was detected but could not be resolved into an **A** or a **G**, one might be tempted to use $L_{j,t}(\mathbf{A}) = L_{j,t}(\mathbf{G}) = \frac{1}{2}$ and $L_{j,t}(\mathbf{C}) = L_{j,t}(\mathbf{T}) = 0$ for the conditional likelihoods at a tip. This would not be correct because the probability of the observation **R** given that the nucleotide is an **A** is 1.0 ($L_{j,t}(\mathbf{A}) = P(R|A) = P(A \cup G|A) = 1$) and similarly if the nucleotide is a **G**. Gaps are treated as ambiguous nucleotides in the substitution models presented here, but see McGuire et al. (2001); Smith et al. (2004) for an alternative treatment.

2.6.2 Combined models, mixture models and time-heterogeneous models

The pruning algorithm has been described here in its simplest form, with a unique substitution model shared at each site and constant throughout the tree. In later chapters, methods that relax this assumption of homogeneity are introduced. The pruning algorithm has to be adapted for that purpose.

Relaxing the constraint of homogeneity in time and across the tree is quite straightforward. One can simply define a specific substitution model for each edge and use those to compute the substitution probabilities in equation 2.15. The pruning algorithm is also easily adapted to change-point models and the process can be made to change inside branches rather than at bifurcation points. In both cases, the process is not stationary anymore and the position of the root has to be properly specified. Ancestral state frequencies at the root should be defined as well since the nucleotide composition is not at equilibrium anymore. From a computational point of view, calculating the likelihood with a time-heterogeneous model is not more expensive but the global evolutionary model usually becomes parameter-rich and fitting the free parameters to the data requires more processing time. Time-heterogeneous methods are discussed in chapter 4.

Relaxing the constraint of homogeneity across sites is not necessarily challenging either but at least two different methods can be used. There are cases where evolutionary patterns are known to be different and sites can simply be partitioned into different categories before the analysis. It is now quite common to concatenate genes to perform a phylogenetic analysis and using a different substitution model for each gene is usually justified. There are also known cases of heterogeneity within a gene. For instance, when protein-coding DNA sequences are studied, one can define a partition in three sets corresponding to the three codon positions. It is also possible to use multiple substitution models to accommodate differences in the protein secondary structure (e.g., alpha-helix, beta-sheet). Obviously, when RNA genes are studied with the base-pair models implemented in **PHASE**, loop and stem regions also have to be partitioned beforehand so that a standard 4-state substitution model can be assigned to unpaired nucleotides and a doublet substitution model can be used with pairs. Combined models that use different substitution models for different blocks of a partition have been studied and are commonly used (Yang, 1996b; Pupko et al., 2002). The overall likelihood is still the product of the likelihood at each site (equation 2.11) and one just has to use the appropriate model for each site depending on the partitioning. Larger partitions will naturally dominate the final likelihood score (Seo et al., 2005), but this is not necessarily an issue.

Substitution models used in different blocks of a partition do not need to be completely independent and **PHASE** has been modified recently towards a more flexible modelling freedom and allows these different substitution models to share some parameters (Yang, 1996b). Different blocks can also use different sets of branch lengths but this functionality is not yet implemented in **PHASE**. Instead, the proportional branch lengths model is used: it is assumed that branch lengths for different classes are the same, up to a scaling factor. This is admittedly a limitation but one has to concede that previous studies did not always confirm the superiority of the complex substitution models that use separate branch lengths (Yang, 1996b; Pupko et al., 2002).

Heterogeneity across sites can also be accounted for using latent class models, also known as mixture models (Pagel and Meade, 2004). Partitioning of the data is not always an option and there are cases where intragenic variability cannot be related to specific DNA segments or a correct partitioning scheme cannot be established with certainty. These models assume that sites evolve according to an unobserved process chosen among a finite set of substitution models and the likelihood is computed by integrating over all the possible substitution processes. As Felsenstein (1981) already pointed out, the expression of the likelihood at a site becomes:

$$\begin{aligned} L_j =& P(X_j|\tau,\nu,\theta) \\ =& \sum_{c=1}^{C} P(X_j|\tau,\nu,\theta_c)p(c) \quad , \end{aligned} \tag{2.18}$$

where C is the number of substitution process (or number of categories), θ_c the subset of parameters in θ that completely defines the evolutionary process for the category c and $p(c)$ the proportion of sites that are assumed to belong to that category. Mixture models are discussed again in chapter 5 and are not described further here. Note that modelling heterogeneity across sites with latent class models introduces more parameters and also increases the computational burden of the likelihood computation (which is proportional to C).

Chapter 3

Bayesian Phylogenetics

In this chapter, the Bayesian approach to phylogenetic inference is discussed as an alternative to the standard Maximum Likelihood (ML) approach. The results of the ML method are solely based on the likelihood, which is the probability of observing the data given the hypothesis (an evolutionary model in our case). The Bayesian method combines the likelihood with the prior for parameters, which is the uncertainty about their true values before the data are known, to generate posterior distribution of parameters upon which the inference is based. Markov chain Monte Carlo methods are discussed at length. These allow for the estimation of species phylogenies with complex models of sequence evolution and are responsible for the current popularity of the Bayesian method in this field of research.

3.1 Introduction

In the previous chapter, an evolutionary model $\mathcal{M}$, with a set of parameters (topology τ, branch lengths ν, substitution parameters θ), has been constructed to model some data X. The Maximum Likelihood (ML) approach considers that the parameter values ($\hat{\tau}$,$\hat{\nu}$,$\hat{\theta}$) that maximize the likelihood function $P(X|\tau, \nu, \theta, \mathcal{M})$ are the best estimates for the parameters of the model. A different approach is followed in Bayesian inference. The problem is still resolved by the formulation of an adequate model that can explain the observed data but the parameters of this model are considered as random variables throughout the analysis.

At the first level of inference, the model $\mathcal{M}$ is assumed to be correct and we aim at inferring the parameters of this model. Bayesian inference starts with a probability distribution that expresses our prior knowledge/belief about the uncertain quantities before the data has arrived. This prior belief is then altered in light of the data using Bayes' rule:

$$\overbrace{p(\tau,\nu,\theta|X,\mathcal{M})}^{posterior} = \frac{\overbrace{P(X|\tau,\nu,\theta,\mathcal{M})}^{likelihood} \times \overbrace{p(\tau,\nu,\theta|\mathcal{M})}^{prior}}{\underbrace{P(X|\mathcal{M})}_{evidence}} \quad . \tag{3.1}$$

The *prior* is combined with the likelihood function to incorporate the new information provided by the data. This gives the conditional probability distribution of the random variables given the data, or *posterior*, upon which Bayesian inference is based. The likelihood is the probability of the data assuming that the model parameters are true. This quantity is not as easily interpretable as the posterior distribution, which is the intuitive quantity a biologist is interested in. Indeed, the posterior probability distribution is the probability of the parameters (tree topology included) given the data. Many important evolutionary questions can easily be answered by marginalization of this quantity (monophyly of a set of species, estimation of ancestral sequences, estimation of divergence times, etc).

Although Bayes' theorem provides us with a rational method to update our beliefs with the arrival of new observations, the choice of a particular prior is problematic and the Bayesian framework is still controversial in statistics. The first issue is that universality is required for a prior to be objective. Two researchers could reach opposite conclusions with the same data if their initial prior is different. Perhaps more importantly, it is expected that equivalent beliefs are always translated into equivalent mathematical expressions. An important philosophical issue is whether numbers can truly be used to reflect beliefs (*e.g.*, can we assign a probability to the existence of extraterrestrial life and, assuming such a number exists, how can we decide its value?). These problems are largely beyond the scope of this thesis since they are epistemological controversies centered on the scientific method. To summarize the problem, Likelihoodists argue that the whole Bayesian framework is subjective because the prior used is subjective in the first place.

In practice, posterior distributions are influenced less and less by the prior and more and more by the likelihood as new data are taken into account. The particular choice of a prior is consequently of limited consequence with enough data. The main advantage of the Bayesian framework over the ML framework is that results incorporate uncertainty over the parameters of the model whereas ML estimates are only point estimates (typically accompanied with a measure of error). Considered altogether with the ability of the Bayesian framework to deal with much more complex models, the inconvenience associated with the use of a prior is not such a big price to pay.

The *evidence* in equation (3.1) is seen as a convenient normalizing constant for the moment but it will be discussed in chapter 4, when the issue of model selection arises. The evidence is the marginal probability of the data and it can be calculated by integrating over all possible parameter values. In the phylogenetic case, the topology is a part of the evolutionary model and computing the evidence also involves summing over all N_T possible topologies. For our purpose, equation (3.1) can consequently be written as:

$$p(\tau,\nu,\theta|X,\mathcal{M}) = \frac{P(X|\tau,\nu,\theta,\mathcal{M}) \times p(\tau,\nu,\theta|\mathcal{M})}{\sum_{i=1}^{N_T}\int_{\nu_i}\int_{\theta} P(X|\tau_i,\nu_i,\theta,\mathcal{M})p(\tau_i,\nu_i,\theta|\mathcal{M})} \quad , \tag{3.2}$$

where τ_i is the ith topology, and ν_i an associated set of branch lengths. The denominator of equation (3.2) is analytically intractable, and, even for small-sized problems, it is usually quite difficult to compute the evidence. Fortunately, numerical methods can be used to approximate the posterior distribution without computation of the evidence. Such a method is discussed in the next section.

3.2 Markov chain Monte Carlo methods

Markov chain Monte Carlo (MCMC) methods are powerful numerical integration methods. Although they cannot be used to compute the posterior probability in a straightforward manner, they can generate large samples from this distribution without explicit integration over all possible topologies and continuous parameters. The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), which generates a Markov chain whose states are the parameters of the complete evolutionary model $\phi = \{\tau, \nu, \theta\}$, is described here.

Algorithm 3.1: The Metropolis-Hastings algorithm

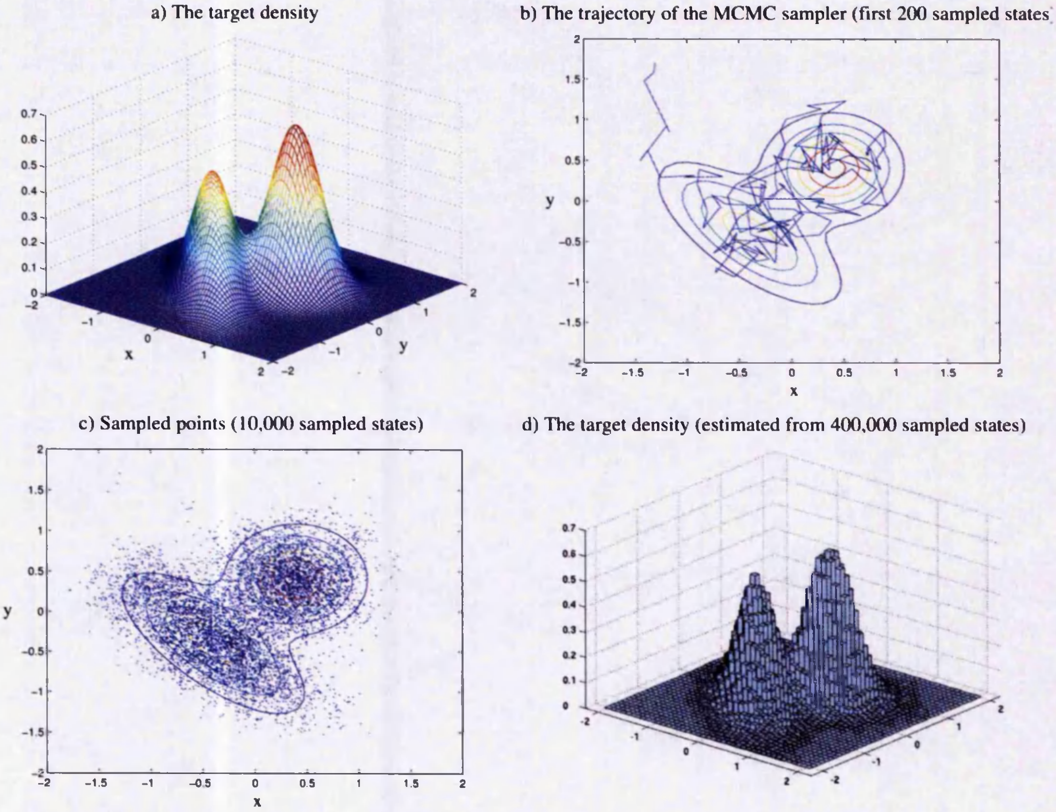
```
/*Initialization*/
$n \leftarrow 1$ ;
Choose a random initial state $\phi_1 = \{\tau_1, \nu_1, \theta_1\}$ ;

/*Sampling*/
repeat
    Propose a new state $\phi' = \{\tau', \nu', \theta'\}$ from the current state $\phi_n$ using a
    proposal mechanism $f(\phi'|\phi_n)$ ;
    Compute $A = \min\left(1, \frac{p(\phi'|X)f(\phi_n|\phi')}{p(\phi_n|X)f(\phi'|\phi_n)}\right)$ ;
    $p \leftarrow$ random($\mathcal{U}(0,1)$) ;     /*draw a random number from $\mathcal{U}(0,1)$*/
    if ( $p < A$ )     /*accept new state with probability $A$*/
    then
        $\phi_{n+1} \leftarrow \phi'$ ;
    else
        $\phi_{n+1} \leftarrow \phi_n$ ;     /* or repeat the current state*/
    end
    Add the state $\phi_{n+1} = \{\tau_{n+1}, \nu_{n+1}, \theta_{n+1}\}$ to the sample;
    $n \leftarrow n + 1$ ;
until enough samples have been collected;
```

The Metropolis-Hastings algorithm (3.1) is a two stage procedure. In the first stage, a new state ϕ' is proposed by modification of the current state ϕ_n. In the second stage, the new state is either accepted or rejected depending on the posterior probabilities $p(\phi_n|X)$ and $p(\phi'|X)$. The acceptance and rejection probabilities at step n are

$$\begin{aligned} P(\phi_{n+1} = \phi'|\phi_n) &= \min\left(1, \frac{p(\phi'|X)f(\phi_n|\phi')}{p(\phi_n|X)f(\phi'|\phi_n)}\right) \\ P(\phi_{n+1} = \phi_n|\phi_n) &= 1 - \min\left(1, \frac{p(\phi'|X)f(\phi_n|\phi')}{p(\phi_n|X)f(\phi'|\phi_n)}\right) \quad , \end{aligned} \tag{3.3}$$

and these define the Markov chain completely. The Hastings ratio $\frac{f(\phi_n|\phi')}{f(\phi'|\phi_n)}$ depends on the proposal mechanisms used and can intuitively be understood as a correction term for when the proposals are not well balanced, *e.g.*, if the proposal mechanism is biased and tends to propose $\phi \rightarrow \phi'$ more often than $\phi' \rightarrow \phi$. One can see from equation (3.3) that the next state is the same as the current state if the proposal is rejected, in which case the state is repeated in the sample. One can also note that a step in the "correct" direction, with $p(\phi'|X) > p(\phi_n|X)$, is always accepted if proposals are well balanced and the Hastings ratio is 1.0.

This Markov chain converges to an equilibrium under quite weak conditions and, once the convergence is reached, states are distributed according to the posterior probability density $p(\phi|X)$. A necessary and sufficient condition for the MCMC algorithm to work is that the chain is *ergodic*: for any couple of states, there must exist a finite number of proposals to travel from one to the other. In other words, all states must be reachable from any initial conditions. Figure 3.1 shows an example of the Metropolis-Hastings algorithm applied to a simple two dimensional density.



a) The target density

b) The trajectory of the MCMC sampler (first 200 sampled states

c) Sampled points (10,000 sampled states)

d) The target density (estimated from 400,000 sampled states)

Figure 3.1: Integration with MCMC methods. the Metropolis-Hastings algorithm that samples from a target distribution is illustrated: a mixture of two Gaussians, shown in a), which is known in this example. b) Each iteration, the MCMC sampler chooses a direction at random and attempts a step in that direction. The step size is chosen from a normal distribution ($\mu = 0, \sigma = .25$) and it can be negative. Steps are small and the chain moves slowly with a high acceptance rate. c) The MCMC sampler is run to produce a large sample. d) The sampled states can be used to reconstruct the initial distribution (a histogram was produced here).

At first glance, there is no clear reason why the transition probabilities in equation (3.3) should be easier to compute than the posterior probability in equation (3.2). However, the denominator in equation (3.2) cancels out when computing the ratio of posterior probabilities in equation (3.3) and one can bypass the need to compute the evidence term when building the Markov chain. The acceptance ratio can be expressed in terms of a ratio of likelihoods, computed with the pruning algorithm presented in chapter 2, a ratio of user-specified priors and the Hastings term:

$$\begin{aligned} P(\phi_{n+1} = \phi'|\phi_n) &= \min\left(1, \frac{P(X|\phi')p(\phi')f(\phi_n|\phi')}{P(X|\phi_n)p(\phi_n)f(\phi'|\phi_n)}\right) \\ P(\phi_{n+1} = \phi_n|\phi_n) &= 1 - \min\left(1, \frac{P(X|\phi')p(\phi')f(\phi_n|\phi')}{P(X|\phi_n)p(\phi_n)f(\phi'|\phi_n)}\right) \quad . \end{aligned} \tag{3.4}$$

Once the sample is generated, one can compute the posterior probability of any identifiable phylogenetic feature of interest by marginalization. For instance, the posterior probability of a specific topology, integrated over the other parameters, is simply given by the fraction of time this topology appears in the sample. Similarly, the posterior probability of a clade being monophyletic is given by the fraction of time this clade appears in the sampled topologies. One can also compute mean posterior estimates for the parameters of the substitution model that accommodate for the uncertainty over tree topologies (see Figure 3.2 for an example of marginalization with the mixture of Gaussian distributions shown in Figure 3.1).

The Metropolis scheme is not the only one available for high dimensional integration. The Gibbs sampler (Geman and Geman, 1984) is a single-component proposal mechanism that modifies a specified parameter using its conditional probability given the other parameters. It is actually a special case of the Metropolis-Hastings algorithm where the ratio of the target density and the Hastings ratio cancel each other out and the proposal is always accepted. The Gibbs sampler is widely used in other fields of research but it did not find a place in phylogenetic inference since there is no analytical form of the required conditional probabilities in most cases.

Dynamic sampling methods have been developed in the field of Physics to study real physical systems. In these methods, particles are associated with their potential energy (which is function of their position) and their kinetic energy.

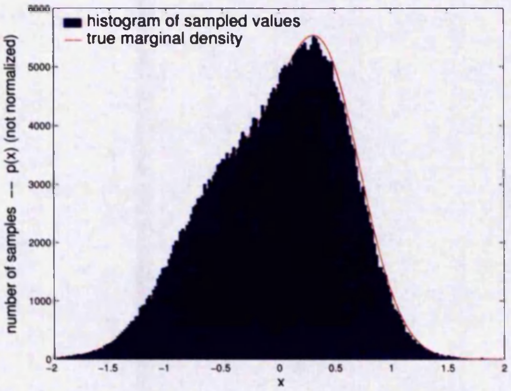


Figure 3.2: Marginalization with MCMC. A high-dimensional model usually has many parameters which are not necessarily of great interest (the so-called "nuisance" parameters). They can be integrated out leading to probability distributions of parameters that are of interest. Using the example previously shown in Figure 3.1, it is assumed here that the y parameter is not as important as the x parameter and a histogram of the x coordinates is produced from the sampled states to approximate $p(x) = \int_y p(x, y)dy$. The histogram is compared with the real density $p(x)$ which is known in this simulated example.

Moves in the phase space are designed to keep the total energy constant. These methods avoid the "random walk" behavior of the Metropolis algorithm and can be faster for some problems. Based on these dynamical methods which can potentially perform large steps while maintaining a high acceptance rate, Hybrid Monte Carlo methods (Duane et al., 1987) were consequently developed in an attempt to improve the mixing of the standard Metropolis algorithm. However, to the best of our knowledge, these hybrid methods have surprisingly not been adapted to the phylogenetic problem. These methods could potentially be used to improve dramatically the mixing of branch lengths for instance. However, since they require the computation of many partial derivatives ($\{\frac{\delta p(\phi|X)}{\delta \nu_i}\}$), it is possible that the increased computational cost outweighs the benefits. Obviously, these methods are restricted to continuous-valued parameters and are not appropriate for discrete parameters such as tree topologies.

3.3 Priors

As mentioned in the introduction of this chapter, the concept of prior is not easy to grasp and is the cause of many controversies. The prior can be based on solid information, easily expressed into mathematical terms, in which case the application of Bayes theorem is uncontroversial. However, in most problems, the prior is of a subjective or arbitrary nature. Indeed, it is rarely possible to translate properly *a priori* beliefs into a mathematical prior. For these reasons, some people remain uncomfortable with Bayesian methods and, most of the time, an "uninformative" prior is chosen in an attempt to remain objective. The prior can also be chosen for mathematical convenience when opportunities to lighten the computational load arise (*e.g.*, conjugate priors). However, note that a real Bayesian would probably describe these attempts as misguided. All priors contain some information and their mathematical expressions should ideally be the best representation of what a researcher thinks before he/she gathers the data.

Phylogeneticists can probably ignore this debate because results are usually not much influenced by the prior when enough data are provided (see Figure 3.3 and Huelsenbeck et al., 2002). Nevertheless, recent works have clearly demonstrated that one should remain cautious, and practical issues encountered with problematic priors are highlighted in this section.

There is generally no strong evidence for a particular prior distribution for the parameters of our evolutionary model. Like most Bayesian phylogenetic inference software, **PHASE** uses a simple factorized prior:

$$p(\phi) = p(\theta) \times p(\nu|\tau) \times P(\tau) \quad . \tag{3.5}$$

$P(\tau)$ is the prior for a given tree topology. By default, **PHASE** uses a simple uniform prior and each topology is given equal prior probability, *i.e.*, $\forall\tau, P(\tau) = 1/N_T$. However, note that complex priors derived from a birth-death generative model for the speciation process, could have been implemented and used with ultrametric trees (see Yang and Rannala, 1997). It is difficult for the biologist to specify an *a priori* knowledge on the space of topologies using probabilities but he might be able to say with complete certainty whether a clade is monophyletic or not. Such knowledge can be inserted before an analysis and it is possible to enforce topological constraints with the **PHASE** software. When performing

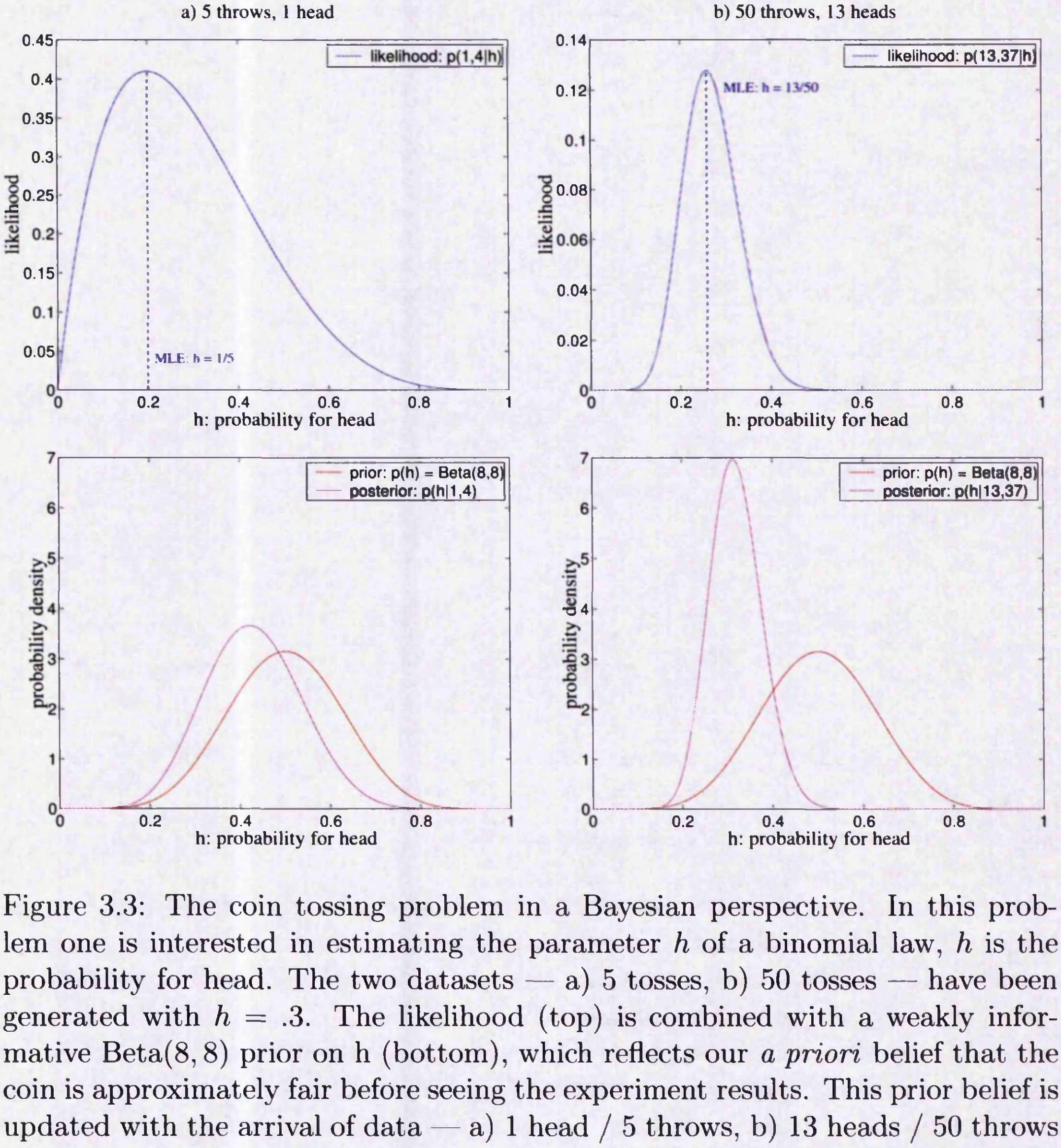


Figure 3.3: The coin tossing problem in a Bayesian perspective. In this problem one is interested in estimating the parameter h of a binomial law, h is the probability for head. The two datasets — a) 5 tosses, b) 50 tosses — have been generated with $h = .3$. The likelihood (top) is combined with a weakly informative Beta(8, 8) prior on h (bottom), which reflects our *a priori* belief that the coin is approximately fair before seeing the experiment results. This prior belief is updated with the arrival of data — a) 1 head / 5 throws, b) 13 heads / 50 throws — to produce the posterior (bottom). Note that ML estimates are exactly the observed fraction of heads. Note also that the likelihood dominates the posterior with the larger dataset.

a Bayesian analysis, the user can specify a list of monophyletic clades or even assume that the complete branching pattern is known[1].

$p(\nu|\tau)$ is the prior for the set of branch lengths associated with the topology. Originally, **PHASE** used a uniform "uninformative" prior and all possible sets of branch lengths were assumed to be equally probable provided that all lengths

[1] This functionality is also available with ML heuristic methods that search for the best tree.

were positive values bounded by a user defined upper limit. Unfortunately, there is really no such thing as an uninformative prior and it has long been appreciated that using flat priors on different parameterizations would produce different posterior probabilities. Unlike the ML method, Bayesian inference does not have the convenient property of being scale-invariant (Felsenstein, 2004). As previously pointed out, the particular choice of a prior is of little importance when enough data are used, unless the prior is unreasonable. It turns out that a uniform prior on branch lengths is far from being even a "vague" prior because this prior attaches a high probability to long branches which, in turn, biases the Bayesian posterior probability of monophyletic clades upwards (Yang and Rannala, 2005). Other priors were consequently made available to the user (exponential as in **MrBayes**, gamma, etc). The default Exp(10) prior is used for most of the inferences presented in this thesis. 10 is the scaling parameter of the exponential law which makes the mean branch length equal to 0.1. This prior seems to be biased towards smaller values and one might reasonably argue that this prior is too arbitrary. A simple solution is to try multiple values for the scaling parameter to check that it does not affect the results. Alternatively, a better solution is to create a *hierarchical* Bayesian model. Under this scheme, the scaling parameter of the exponential distribution becomes a "hyper-parameter" of the model. The hyper-parameter must be given a "hyper-prior" distribution (*e.g.*, $\lambda \sim \text{Exp}(1)$ or $\lambda \sim \mathcal{U}(0, 50)$) and is estimated like a standard parameter during the MCMC run. It is possible to create such hierarchical models with **PHASE**. This functionality was only used when enough computational power was available since it was not found to have a significant impact on the results.

Such priors on branch lengths cannot be used with an ultrametric tree. In such a case, the prior on branch lengths is replaced with a prior on the height h of the tree:

$$p(\nu|\tau) = p(\nu|h,\tau)p(h|\tau) \quad , \tag{3.6}$$

The user can choose the prior on the height $p(h|\tau)$ (uniform, exponential, etc) and all valid sets of branch lengths that match the given height have equal probabilities.

Finally, $p(\theta)$ is the prior on the parameters of the substitution model described in the previous chapter. Once again, a factorized prior was chosen and $p(\theta)$ is the product of the priors on each parameter. A Dirichlet distribution is used for the state frequency vector $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$:

$$p(\Pi) = \frac{\Gamma(\sum_i p_i)}{\prod_i \Gamma(p_i)} \prod_i \pi_i^{p_i - 1} \delta(\sum_i \pi_i - 1) \quad , \tag{3.7}$$

where δ is the Kronecker function and $\{p_i\}$ are the parameters of the Dirichlet distribution. The user can specify these parameters and, in this thesis, the default flat Dirichlet prior, $\forall i, p_i = 1$, is used. This prior attaches a uniform probability density to all sets of frequency parameters that sum up to one. If this prior is judged arbitrary, a hierarchical scheme can also be used where $p_0 = \sum_i p_i$ becomes a hyper-parameter of the model that controls the variance of the Dirichlet prior. In such a case, the user still has to specify the center of the Dirichlet distribution $\{\frac{p_i}{p_0}\}$ and to decide on a hyper-prior on p_0.

The situation is slightly more complex for the exchangeability parameters. As mentioned in section 2.5, it is necessary to impose an extra constraint on the exchangeability parameters since they can only be identified up to a scaling factor. Most ML phylogenetic programs use one of these parameters as a reference and set its value to 1.0. The other exchangeability parameters are then given relatively to this reference. Hence exchangeability parameters are often called rate ratios. In chapter 5, this older parameterization is used and a uniform prior, between 0.0 and an arbitrary upper-bound of 200.0, is used on individual rate ratios. The reference rate is $\rho_{A \leftrightarrow G}$ when a DNA model is used and is usually $\rho_{AU \leftrightarrow GC}$ with base-pair models (but $\rho_{AU \leftrightarrow GU}$ is used as a reference when double-substitutions are not allowed). However, experience accumulated over the last three years has shown that this prior was problematic when performing a Bayesian inference with RNA substitution models. With a limited amount of data, *e.g.*, small sequences or few species, rate ratios were seen to reach the upper-bound imposed by the prior. As was recently pointed out, the uniform prior on rate ratios produces biased parameter estimates because it puts more weights on higher values (Zwickl and Holder, 2004). The problem gets worse when using a high upper boundary or an unbounded uniform prior in a misguided desire of objectivity. In chapter 5, this issue is of no consequence since the likelihood dominates the prior in determining the posterior distribution of the exchangeability parameters. Nevertheless, a parameter-rich model is used in chapter 4 and this issue would have been a more significant concern. A newer parameterization was consequently used and the constraint on exchangeability parameters was changed to be $\sum_{i=1}^{nb_{exch}} \rho_i = 1$. The vector of exchangeability parameters is now similar to the state frequency vector and the various Dirichlet priors described above can be used. As suggested in Zwickl and Holder (2004), the default Dirichlet parameters, $\forall i, p_i = 0.5$, are used in this thesis.

In chapter 5, mixture models that allow for rate heterogeneity across sites are described. Default uniform priors are used on the extra parameters introduced by these techniques, *e.g.*, $\mathcal{U}(0.0, 1.0)$ for the proportion of invariant sites, $\mathcal{U}(0.0, 50.0)$ for the gamma shape parameter that controls the extent of rate heterogeneity. However, **PHASE** allows for flexible modelling and different priors could have been chosen for these parameters.

Finally, some MCMC runs presented in this thesis are performed with a combined substitution model. Real RNA sequences are partitioned before an analysis and the average substitution rate in loops and stems is not assumed to be the same. As mentioned in section 2.6, **PHASE** is using the proportional branch lengths model. In practice, this translates into each substitution model having a different average substitution rate that remains constant over the whole tree, plus one constraint that gives a meaning to the branch lengths (see section 2.5). In chapter 5, the average substitution rate of the loop partition is fixed to 1.0 substitution per site and per unit of branch length. The average substitution rate of the stem partition is a free parameter (c) of the evolutionary model and a uniform prior $\mathcal{U}(0.0, 200.0)$ is attached to it. This prior is not symmetric and implies a higher probability for the stems to evolve faster. As was the case with the uniform prior on rate ratios, this can potentially lead to biasing effects during real inference (PG Higgs, personal communication). For the complex substitution model presented in chapter 4, the constraint was changed and the sum of the average substitution rates is fixed to be equal to the number of substitution models, *e.g.*, $c_1 + c_2 = 2.0$ when different Markov processes are used for loops and stems. For the MCMC runs presented in this thesis, a flat Dirichlet prior is attached to $\{c_i / \sum_i c_i\}$.

3.4 Proposal distributions

The transition proposals used to move through the state space are of crucial importance for the effectiveness of the MCMC sampler. One has to balance the desire to move globally and efficiently through the parameter space with the need to make computationally feasible and reversible moves with a known Hastings ratio $f(\phi_n|\phi')/f(\phi'|\phi_n)$. Distant moves in the space of parameters may allow us to traverse the space of parameters quickly but such moves will usually be associated with low acceptance rates that deteriorate the performance of the sampler. On the other hand, timid steps with a high acceptance rate do not allow for a complete exploration of the highly probable areas. Proposal algorithms have consequently to be designed with great care to ensure proper mixing.

It would be computationally difficult, and useless, to update all the parameters at once when complex models are used. The state space can hopefully be divided, a technique known as *blocking*, and its components can be updated separately. This is not always a good strategy and components that are highly correlated should ideally be grouped together in blocks to be updated simultaneously by a proposal distribution that takes the correlation into account (Yang, 2005). Nevertheless, the parameters of the phylogenetic model are usually weakly correlated[2] and the different components can be considered independently. To make a new proposal, one of the blocks is randomly selected for update at each step and the user can tune the different update probabilities to improve the mixing behavior. For example, it is advisable to increase the probability of updating the topology when lots of species are used and the number of possible phylogenies grows.

It is worth pointing out at this point that *any* proposal distribution can (and will) give valid results if it is run for an appropriate amount of time. The only requirement being, as was previously mentioned, that a path exists between any two points of the state space. Consequently, the choice of a particular proposal is ultimately of limited consequence but might have a huge impact on the simulation time. It has been said that the choice of good proposal distributions involves "the burning of incense, casting of chicken bones, use of magical incantations, and invoking the opinions of more prestigious colleagues" (Felsenstein, 2004). Indeed, various proposal mechanisms have already been proposed and used for the phylogenetic problem but their efficiency has not been extensively studied yet and the particular choice of a proposal is a matter of experience at the moment. However, such a statement will probably spur phylogeneticists to close this gap

[2] The proportion of invariant sites and the gamma shape parameter described in chapter 5 are a notable exception.

with solid scientific studies in the near future.

3.4.1 Proposals for continuous parameters

For the purpose of MCMC proposals, the parameters of the substitution model can be organized into two classes: independent scalars and vectors of dependant values. Independent values are updated independently but the parameters of a vector are updated together in a single MCMC step. To update the parameters of a vector, the proposal scheme of Larget and Simon (1999) is adopted. This proposal mechanism is used with the state frequency vectors but also with the vector of exchangeability parameters and the vector of average substitution rates when the constraint used is that they sum up to 1.0. A new set of values is proposed using a Dirichlet distribution centered at the current values:

$$f(\Pi'|\Pi) = \frac{\Gamma(p_0)}{\prod_i \Gamma(p_0 \pi_i)} \prod_i {\pi'_i}^{p_0 \pi_i - 1} \delta\left(\textstyle\sum_i \pi'_i - 1\right) \quad , \tag{3.8}$$

where $\Pi = \{\pi_i\}$ is the current vector, $\Pi' = \{\pi'_i\}$ is the proposed vector and p_0 is a value that controls the variance of this distribution. For large values of p_0 the distribution is tight and the new vector is more likely to be closer to the current set of values. For implementation purpose, one should mention that one can sample from this Dirichlet distribution by combining samples from gamma distributions with parameters $\{p_0 \pi_i\}$, and normalizing those. One should also note that the Hastings ratio of this proposal mechanism is not trivial and is equal to $f(\Pi|\Pi')/f(\Pi'|\Pi)$. The choice of p_0 is crucial for the mixing behaviour. In the earliest version of **PHASE**, p_0 was user-specified, but this value is now automatically tuned during the burn-in period to reach a reasonable acceptance rate (between 20% and 25% by default). The acceptance rate is computed every 200 iterations and p_0 is respectively multiplied/divided by a tuning factor if the acceptance rate is found to be lower/higher than the chosen boundaries so that the step size is reduced/increased. When the acceptance rate starts oscillating and switches between values higher than 25% and lower than 20% this tuning factor is gradually reduced so that changes of p_0 become smaller. This complex mechanism is naturally turned off when the sampling begins because it would contravene the MCMC principles.

A sliding window mechanism is used to update independent parameters. This proposal is used, for instance, with the proportion of invariant sites and the gamma shape parameter described in chapter 5. It is also used with exchangeability parameters and average substitution rates when the "rate ratios parameterization" is used and one of these parameters is constrained to be equal to 1.0. For independent parameters, the new value is drawn from a Gaussian distribution centered on the current value:

$$f(x'|x) = \frac{1}{\sigma_x\sqrt{2\pi}} \times e^{-\frac{(x'-x)^2}{2\sigma_x^2}} \quad . \tag{3.9}$$

Each independent parameter of the substitution model has a corresponding standard deviation σ_x which is modified during the burn-in to reach a reasonable acceptance rate. A similar tuning mechanism has been described above for vectors of parameters and its principles are the same. The main difference is that the standard deviation has to be lowered to reduce the step size when the acceptance rate is not large enough. Since $f(x'|x) = f(x|x')$, the Hastings ratio for such proposals is 1.0.

One issue arising with the use of a normal distribution is that it is possible to propose values outside the allowed interval whenever a parameter has an upper and/or lower bound. Theorically, the issue is not critical because such moves would fall outside the area of the state space allowed by the prior distribution and would be automatically rejected. Practically, such doomed proposals are wasting computational resources and should be avoided whenever possible. Reflecting boundaries are consequently used to ensure that the proposed values remain within the allowed range. Quite conveniently, the Hastings ratio is still equal to 1.0 when a reflected Gaussian distribution is used because each reflection from x to x' has a corresponding reflection from x' to x.

$[20\%, 25\%]$ was suggested as an optimal acceptance rate in this section but this is merely an educated guess and any acceptance rate between $[10\%, 60\%]$ would probably be as efficient. Experience has shown that the marginal posterior probability densities of phylogenetic parameters is usually unimodal and one can consult Gelman et al. (1996) for experimental determination of the optimal step size for multivariate normal problems.

3.4.2 Proposals in the space of phylogenies

To search the parameter space, one also has to define moves in the discrete topology space and the associated continuous branch lengths space. Jow et al. (2002) designed the proposals used with unrooted topologies in **PHASE** and these proposals have been reimplemented to optimize the likelihood computation and to cope with new functionalities, *e.g.*, the definition of monophyletic clades that constrain the space of possible topologies. Algorithms developed for unrooted trees are not compatible with ultrametric or rooted trees and therefore two additional sets of moves have also been designed.

The first set of moves is used with rooted ultrametric trees, when a global molecular clock is assumed. Proposals that preserve the distance from the root to the tips were required and are described along with the standard proposals for unrooted trees in this section. The second set of moves is used with the time-heterogeneous substitution models developed in chapter 4. With non-reversible models, the likelihood depends on the position of the root and it is consequently natural to use rooted topologies with them. Standard proposals for unrooted trees have been slightly modified to account for this peculiarity but they are not introduced until the next chapter, where their use can be properly described.

MCMC algorithms are rejection-based techniques and, for computational reasons, one should not discard past computations too quickly. If the new state is rejected, partial likelihood arrays (see section 2.6) of the old phylogeny are still valid and might still be useful to compute the likelihood of the next new state. When describing proposals in the space of topology, some nodes are said to be "invalidated", meaning that the partial likelihoods of these nodes has to be computed when calculating the likelihood of the new state whereas the computations previously performed at other nodes are still valid and can be reused. The partial likelihoods of invalidated nodes are safely kept until the end of the iteration. They are discarded if the proposal is accepted but, if it is not, they are restored when backtracking to the old state[3]. Similar optimizations are also used when proposing new values for the substitution model parameters. With standard substitution models, all nodes have to be invalidated when such a proposal is attempted. Nevertheless, when multiple substitution models are used in

[3] Howsun Jow's implementation recognized that fact but his backtrack mechanism involved the copy of large chunks of memory. The new implementation is more elegant and swifter since it is only permuting pointers to some allocated memory space.

a combined analysis of partitioned data, some proposals might only affect the limited number of sites belonging to the affected block. In such a case, one only has to invalidate the columns of the alignments that were affected by the change.

Topology proposals

Local and global moves are used to propose new topologies. Candidate trees proposed by local moves are "closer" to the initial tree and consequently have a better acceptance rate. Nevertheless, it was found that these local moves are insufficient for proper exploration of the space since the chain of "local intermediates" between two distant and highly probable topologies can sometimes contain very unlikely trees. Additional proposals, which allow for larger changes in the topology, have consequently been developed and are used conjointly with the local proposal mechanisms.

The Nearest Neighbor Interchange (NNI) is our local proposal and is shown in Figure 3.4 for unrooted trees. A random internal branch is chosen and two subtrees or leaves linked to opposite sides of that branch are swapped. If one side of the branch was defined as a monophyletic cluster (*e.g.*, (A,B) in the figure), the NNI is not allowed and another branch is chosen.

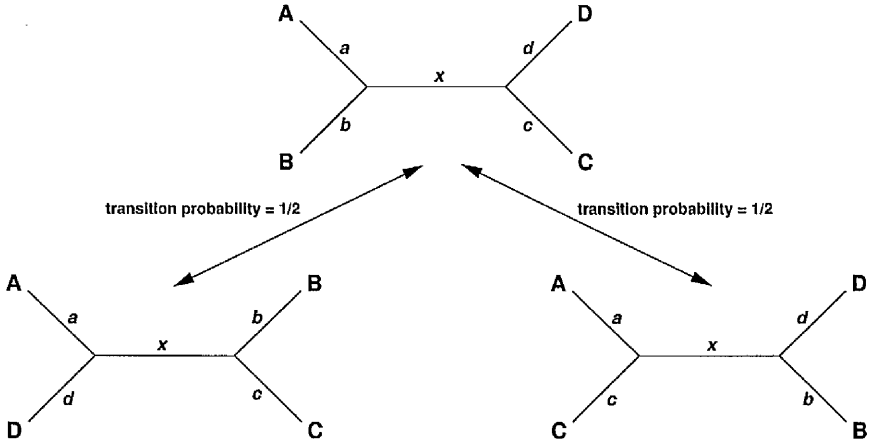


Figure 3.4: Nearest Neighbor Interchange for unrooted trees. Lowercase letters represent branch lengths.

The NNI proposal for ultrametric trees is slightly different. An internal node is chosen at random among those which are not directly linked to two leaves and the branch leading to its closest child is chosen to define the four subtrees **A**,

B, **C** and **D** (see Figure 3.5). Another branch has to be chosen if (C,D) is a monophyletic clade. To maintain ultrametricity, the heights of the nodes are not modified and branch lengths are updated to match this constraint.

The Hastings ratio of the NNI proposal for unrooted and ultrametric trees is 1.0 because the probability of the reverse move is the same. In both cases, the two nodes connected to the internal branch and the nodes between the branch and the root (root included) are invalidated[4].

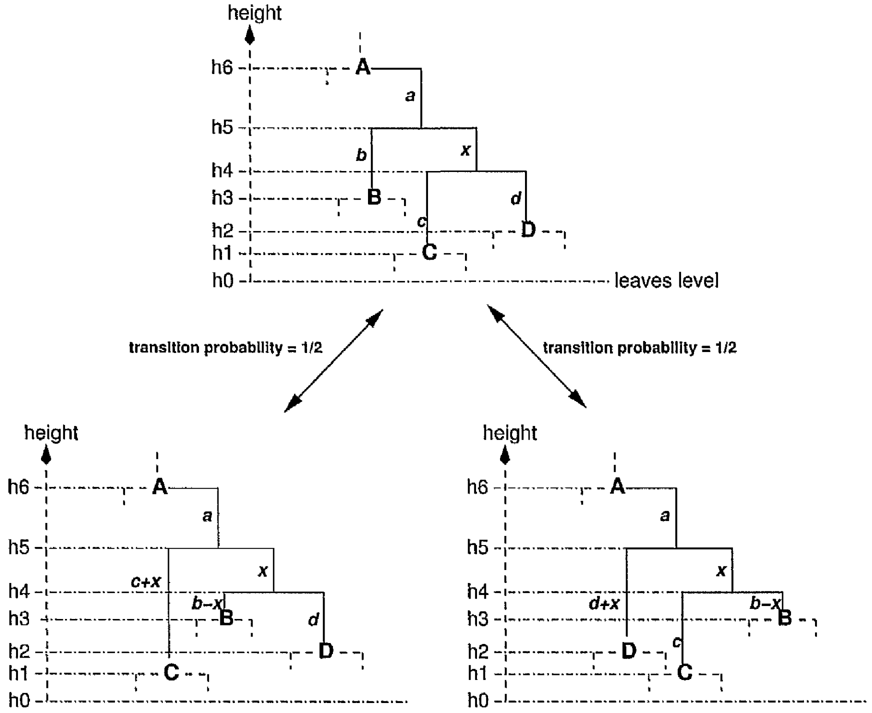


Figure 3.5: Nearest Neighbor Interchange for ultrametric trees. Lowercase letters represent branch lengths.

The Subtree Pruning and Regrafting (SPR) is our long-range move in the space of topologies. The Tree Bisection and Reconnection (TBR) is another well known global move but it is not implemented in **PHASE**. The SPR proposal for unrooted trees, shown in Figure 3.6, can propose a wide range of topologies from the initial state but is characterized by a very low acceptance rate. A branch is chosen at random and is reattached at a random point along a randomly selected branch. The two selected branches cannot be adjacent and two other branches are

[4] It is recalled that a root is arbitrarily chosen to compute the likelihood of unrooted trees with the pruning algorithm.

chosen if the proposal is not compatible with the set of user-defined monophyletic clades. The Hastings ratio for this proposal is not trivial and is related to the particular choice of an attachment point along the selected branch. When a SPR move is attempted, one needs to invalidate the nodes above the detached branch before it is detached as well as the nodes above its insertion point (including the insertion point itself).

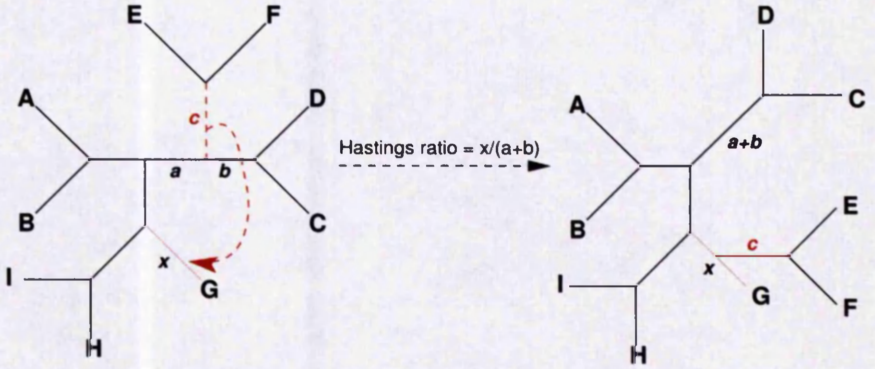


Figure 3.6: Subtree Pruning and Regrafting for unrooted trees. Lowercase letters represent branch lengths.

The SPR proposal for ultrametric tree is quite different and is described in Figure 3.7. For this proposal, a subtree is chosen at random, removed, and then reinserted at a random point above its height. Clade constraints are taken into account and only valid attachment points can be selected. Note that this proposal does not necessarily change the topology because the subtree can be reattached on the original branch. The Hastings ratio for this proposal is 1.0 because all possible attachment points have equal probability and this probability would be the same for the reverse move. One has to invalidate the nodes above the detachment and the reattachement points.

Branch lengths proposals and "continuous" topology change

Branch length proposals are designed to change the branch lengths of a phylogeny. As a side effect in **PHASE**, such proposals might also trigger a change in the topology. For unrooted trees, a branch is randomly selected and its new length is drawn from a Gaussian distribution centered at the current value. This proposal mechanism is similar to the sliding window mechanism described previously for the parameters of the substitution model. The standard deviation of this normal

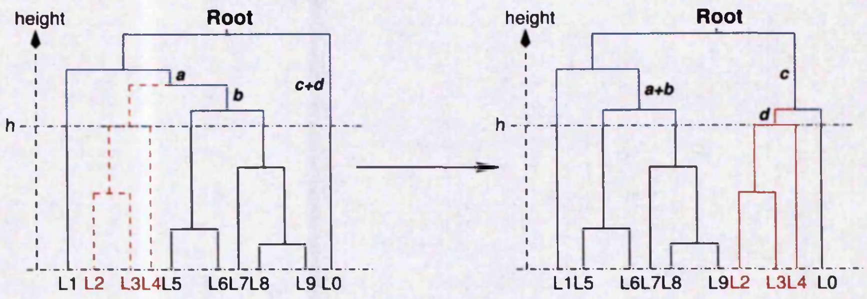


Figure 3.7: Subtree Pruning and Regrafting for ultrametric trees. Lowercase letters represent branch lengths.

distribution is common to all branches and is tuned during the burnin period to reach a reasonable acceptance rate. When a uniform prior is used, proposals above the upper limit are reflected back. Negative values are also reflected back but, if possible, a NNI is simultaneously attempted in such a case. Negative branch lengths are taken as a sign that the topology might not be strongly supported and this proposal may be more likely to be accepted (Jow et al., 2002). This result in a "smooth" change between topologies. The Hasting ratio for this proposal is not affected by this peculiarity and is equal to 1.0.

This proposal method is not compatible with ultrametric trees and a different mechanism was designed for this specific case. This proposal, shown in Figure 3.8, can also prompt a change in the topology. A random internal node is chosen and a new height will be proposed for it. First, a random value β is drawn from a normal distribution centered at 0 and reflected back into $[-D; D]$, where D is the distance between the parent of the chosen node and its closest child. The standard deviation of this distribution is $D \times \gamma$ where γ is a tuning parameter common for all branches and modified during the burnin for optimal mixing. The new height is equal to the old height plus β and the value is reflected to ensure that the node remains below its parent and above its closest child. The proposal triggers an NNI when the height is reflected against the closest child height if it is not a leaf node and if the topological constraints allow it. The Hastings ratio for this proposal is 1.0.

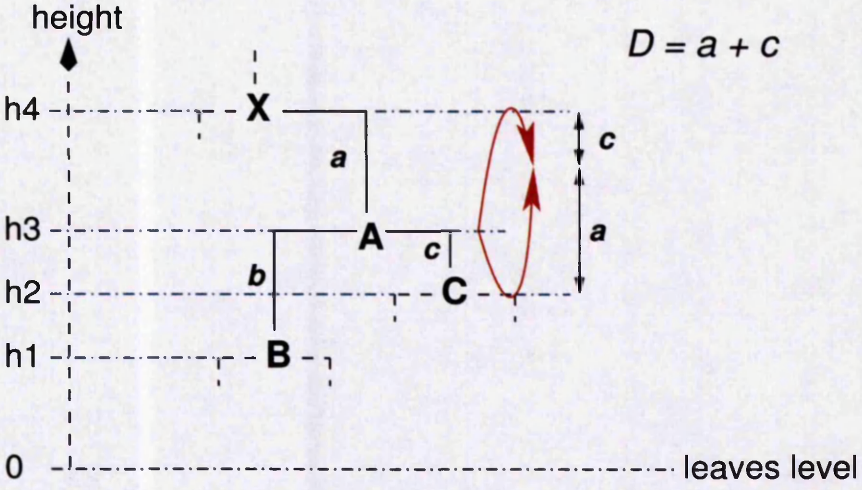


Figure 3.8: Branch proposal for ultrametric trees. Lowercase letters represent branch lengths.
β is drawn from a normal distribution $\mathcal{N}(0, \gamma D)$ and reflected in $[-D, D]$.
If $\beta > a$, $h3' = h4 - (\beta - a)$,
if $a \geq \beta > 0$, $h3' = h3 + \beta$,
if $0 \geq \beta > -c$, $h3' = h3 + \beta$,
if $-c \geq \beta$, $h3' = h2 + (-c - \beta)$ and a local NNI is proposed (see Figure 3.5).

3.5 Practical considerations and issues

Bayesian MCMC programs are typically easy to implement and hard to debug. Bayesian MCMC programs for phylogenetic inference are difficult on both counts. The implementation is difficult and error-prone because a consequent amount of code is devoted to various optimizations and the original algorithms quickly become complex. The temptation to offer many sophisticated evolutionary models in a single software package further adds to the difficulty of the task.

Program validation is even harder. ML programs can be tested with synthetic data because they return a single point estimate. Branch lengths and substitution parameters are supposed to converge to the values that were used to generate the data in the first place. With MCMC methods, the convergence is to a posterior distribution that cannot be computed easily, even with generated data. It is not possible to prove completely the correctness of the algorithms. Nevertheless, the software has successfully passed a wide range of tests and can certainly be considered usable for scientific research.

PHASE can perform both ML and Bayesian inference. Consequently, the pruning algorithm used to compute the likelihood function has been tested in the ML framework and results have been compared with the results of other ML software like **PAML** (Yang, 1997b). Substitution models implemented in **PHASE** were also tested in the process. MCMC algorithms were tested and found to work properly with empty sequences. Indeed, when no data are provided, the posterior distribution is theorically equal to the prior distribution and it was checked that clade posterior probabilities and sampled substitution parameters followed their respective prior distributions. Last but not least, a recovery mechanism was implemented to restore MCMC runs that abort before completion. Since this mechanism is working perfectly, it implies indirectely that all the error-prone optimization techniques that save partial likelihood values between successive iterations are working properly.

Programmers have to be cautious but the user cannot afford to be careless either. Bayesian inference programs for phylogenetic inference cannot be treated as black boxes that output a valid phylogeny when fed with molecular sequences. Outputs of the software have to be processed and checked carefully (Huelsenbeck et al., 2002). The first immediate problem is to determine whether the chain has been run long enough. The likelihood of the visited states should have converged to a stationary distribution but this is not a sufficient condition. The convergence of all the evolutionary parameters should also be checked to spot possible mixing problems.

Since there is also an additional risk of being trapped in a local maximum, a safer approach that can address both problems is used in this thesis. The results presented here are always produced by several MCMC chains started from random initial trees. Convergence is more likely to be assured if four chains, or more, produce similar posterior probabilities for the clades of the phylogeny and the parameters of the substitution model. Moreover, the problem of slow mixing can be detected more easily when multiple chains are run. With multiple chains, it is also possible to attach an estimate of error — the so called Monte Carlo error — to the posterior probabilities produced by the software.

A second issue, which should be appreciated by the user of Bayesian methods, is the sensivity of the results to the chosen prior. As already mentioned, results are usually not affected by a particular prior choice in a typical Bayesian analysis and this problem is not as important as the previous one. Nevertheless, a prior chosen carelessly can become problematic if it is not corrected by providing enough data.

It is consequently good practice to analyse a dataset using different priors to see how robust the results are.

Chapter 4

Heterogeneity in time

There is strong evidence that nucleotide frequencies in nuclear and mitochondrial RNA genes are varying along different lineages. It is thought that this can lead to the recovery of spurious phylogenies because traditional phylogenetic methods assuming homogeneity tend to group together species with similar nucleotide frequencies, regardless of their actual evolutionary relationships. In this chapter, we introduce a nonhomogeneous evolutionary model that accounts for variation of nucleotide and base-pair frequencies over time. This nonhomogeneous and nonreversible model is built with locally homogeneous models by using different substitution matrices on different branches of the tree. The homogeneous base-pair substitution models introduced in previous chapters are already parameter-rich and the available sequence data cannot support the use of an independent substitution matrix on each branch of the phylogeny. A reversible jump Markov chain Monte Carlo technique is consequently used to limit the size of the parameter space while still accommodating for the extent of compositional heterogeneity observed in contemporary sequences.

4.1 Introduction

The nucleotide and base-pair substitution models introduced in chapter 2 are admittedly too simple to model the complex mechanisms involved in the evolution of DNA and RNA sequences adequately. Fortunately, perfect models are not necessarily required for reliable phylogenetic inference, as confirmed by the substantial congruence between phylogenies recovered from different datasets (Hillis, 1995; Sullivan and Swofford, 2001). Nevertheless, it is known that violations of the assumptions of the evolutionary model generally introduce a limited bias in the results of an inference and have an impact on the accuracy of a phylogenetic method.

All existing phylogenetic methods have some shortcomings. For instance, the UPGMA reconstruction algorithm (see section 2.2.3) does not behave well when given distances from a non-clocklike tree, parsimony methods are sensitive to unequal rate variation across sites and model-based ML and Bayesian methods can give very misleading results when some assumptions of the evolutionary model are violated. When evaluating the performance of phylogenetic methods, one is generally interested in three different factors (Huelsenbeck, 1995; Hillis, 1995).

1. Consistency: does the method converge to the correct result (phylogeny and/or substitution parameters) as more data are applied to the problem? This criterion is often advocated to justify the use of model-based ML methods over parsimony-based methods since ML estimators are known to be consistent if certain conditions are met (Felsenstein, 1973, 1978). Nevertheless, note that proof of consistency depends on the generating model and it has repeatedly been shown that the ML method can become statistically inconsistent and converge towards a wrong answer if the evolutionary model used to perform the inference differs from the mechanisms that generated the sequences in the first place (Chang, 1996a; Kolaczkowski and Thornton, 2004)[1]. Consequently, convergence cannot be guaranteed when real sequences are analyzed.
2. Efficiency: how fast is the convergence? In standard statistical problems, ML estimators are the most efficient asymptotically but this property is not verified in phylogenetic inference because the tree topology is not a continuous

[1] The issue can also arise when the parameters of the evolutionary model are not identifiable (Steel et al., 1994)

parameter and each tree defines a separate parameter space (Yang, 1997a). Although this should not be considered as the norm, it has been found that simple and wrong evolutionary models can sometimes outperform complex but correct ones. Since wrong models are usually biased towards specific tree shapes, they are naturally advantaged when the bias is coincidentally towards the true topology (Bruno and Halpern, 1999). In a Bayesian setting, the problem would manifest itself by a higher than desirable support for the correct topology and overconfidence in the results, which is undesirable too (Sullivan and Swofford, 2001).

3. Robustness: how does the method perform when its assumptions are not met? This is perhaps the most important criterion for practicing systematists since assumptions of phylogenetic methods are inevitably violated to an extent with real data. The traditional approach to evaluate robustness is to use simulations to evaluate the behaviour of a method when a limited number of its assumptions are violated in controlled computer experiments.

Standard time-homogeneous substitution models are parameterized with a single composition vector that defines the equilibrium distribution of the substitution process. These models assume that the substitution process is homogeneous and stationary and it is consequently expected that the actual nucleotide composition matches this equilibrium distribution throughout the phylogeny, within an acceptable range of stochastic variation. As a corollary, since it has often been observed that the sequences of the studied taxa can have widely different composition, we have clear evidence that these assumptions are regularly violated with real data sets.

With traditional phylogenetic methods, sequences of similar composition tend to be grouped together irrespective of their real evolutionary relationships. Consequently, the biasing effect of compositional heterogeneity on distance, parsimony and model-based reconstruction methods has long been recognized as a potential issue in the literature (see, *e.g.*, Loomis and Smith, 1990; Olsen and Woese, 1993; Foster and Hickey, 1997; Tarrío et al., 2000; Chang and Campbell, 2000). However, its practical effect on the topological accuracy of reconstructed trees has been downplayed and recent simulations have shown that an extreme amount of heterogeneity is necessary for the compositional bias to have a substantial effect on the reconstructed phylogeny and for traditional methods to become inconsistent (Conant and Lewis, 2001; Rosenberg and Kumar, 2003).

Nevertheless, a contrary viewpoint was expressed by Jermiin et al. (2004) who showed that the biasing effect is strongly dependent on the length of the short internal edges. In practice, and as illustrated in Figure 4.1 with a nuclear LSU RNA dataset containing five bacterial species, the compositional bias can sometimes have visible effects on the reconstructed phylogeny. Although convincing external evidence supports the grouping of the genus *Thermus* with the genus *Deinococcus* (Embley et al., 1993; Foster, 2004), the thermophilic species *Thermus thermophilus* is attracted to the two other thermophilic species when standard homogeneous methods are used with this dataset. As explained later on in this chapter, the G+C content of the SSU and LSU rRNA genes of prokaryote species is correlated to their optimal growth temperature and homogeneous methods presumably failed to recover the correct tree because of the important G+C compositional bias present in the sequences.

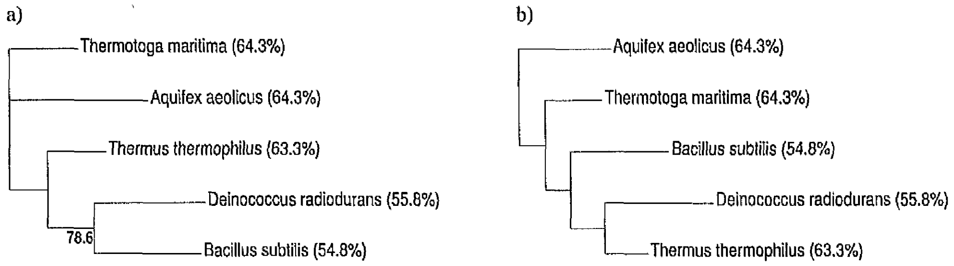

a)
Thermotoga maritima (64.3%)
Aquifex aeolicus (64.3%)
Thermus thermophilus (63.3%)
Deinococcus radiodurans (55.8%)
78.6
Bacillus subtilis (54.8%)
b)
Aquifex aeolicus (64.3%)
Thermotoga maritima (64.3%)
Bacillus subtilis (54.8%)
Deinococcus radiodurans (55.8%)
Thermus thermophilus (63.3%)

Figure 4.1: The compositional bias illustrated with a nuclear LSU dataset: The LSU RNA gene of five bacteria is studied (a) with a standard homogeneous model (TN93) and (b) with the time-heterogeneous method developed in this chapter (TN93+hF: the same model but with heterogeneous frequency parameters). Correlation between opposite sites in RNA stems was not taken into account in this example [3]. The G+C content of each sequence is given after the species name. The number in red in (a) is the Bayesian posterior probability (BPP) for the clade *Deinococcus+Bacillus*. The wrong tree was supported with BPP 78.6% and the correct tree was supported with BPP 21.4%. With the time-heterogeneous model, the correct tree (b) is recovered with BPP 100%. This tree is rooted for reasons that are explained below. Note that other studies have illustrated a similar compositional bias with the same set of species using the SSU RNA gene (Galtier and Gouy, 1995; Foster, 2004).

[3] It turns out that the correct tree would be recovered in both cases if the dataset was partitioned and a base-pair model was used with helices (results not shown).

The development of more realistic models is a necessary step towards a better understanding of the nucleotide substitution process and RNA sequence evolution. In this chapter, a method that accounts for the variation in composition in different lineages is consequently introduced. Since it models the data more accurately, the method is also expected to reduce the systematic error and return more accurate posterior probabilities for the substitution parameters and topology estimates. The method was implemented in **PHASE** and operates in a Bayesian framework using MCMC techniques.

In section 4.2, previous works are reviewed and the basis of the time heterogeneous model is laid out. In section 4.3, the MCMC proposals that modify the placement of composition parameters on the tree are presented. Proposals in the space of topology that are used for unrooted trees (see section 3.4.2) are adapted to handle rooted trees and the presence of extra, discrete, parameters that assign a frequency vector to each branch. The dimension of the parameter space, *i.e.*, the number of composition parameters, is kept constant during these moves. In section 4.4, a reversible jump method is introduced with split and merge moves that are used to increase or decrease the number of composition parameters in the model and to jump between parameter spaces of differing dimensions. In section 4.5, an issue with the prior distributions on new model parameters is highlighted and a hierarchical Bayesian approach that uses a slightly more complex prior with hyperparameters is adopted. These hyperparameters are assigned a hyperprior distribution and are estimated along with other parameters of the model during the inference. In section 4.6, Bayesian simulations are used to evaluate the impact of compositional variation over time, as defined by the time heterogeneous model presented in this chapter, on the corresponding standard homogeneous model. In section 4.7, the time heterogeneous method is applied to real data with a dataset of 40 species spanning the entire tree of life. Results with homogeneous and time-heterogeneous methods are compared.

4.2 A substitution process for compositional heterogeneity

In an attempt to overcome the problem posed by the compositional bias, distance methods that allow for compositional variations over time have been developed

(Lake, 1994; Lockhart et al., 1994; Galtier and Gouy, 1995; Tamura and Kumar, 2002). Alternative model-based ML methods were researched in parallel. Following an early work of Barry and Hartigan (1987), Yang and Roberts (1995) and Galtier and Gouy (1998) proposed some substitution models that assign different frequency parameters to the branches of the phylogeny. Such processes are not at equilibrium and the average nucleotide composition varies along the branches of the tree. Felsenstein's pulley principle (1981) does not apply with these processes that are not time-reversible and the likelihood of such evolutionary models depends on the placement of the root. Consequently, rooted trees have to be considered and it is also necessary to specify the initial state frequencies at the root which can also be considered as a parameter of the substitution model [4].

In Yang and Roberts (1995), the most general time-heterogeneous substitution model, called N2, assigns four nucleotide frequency parameters, *i.e.*, three free parameters, to each branch of the phylogeny. Four extra parameters are used for the ancestral composition at the root of the tree. This model is computationally hardly tractable and Yang and Roberts also proposed a N1 model which uses a common set of frequency parameters for all the internal branches. Yang and Roberts' time-heterogeneous models are based on the homogeneous HKY85 nucleotide substitution model (see chapter 2). Galtier and Gouy (1998) proposed a simplified version of HKY85+N2 by replacing the HKY85 model with the T92 model (Tamura, 1992). In the T92 model, the nucleotide composition is only described by the **G+C** content and is obtained by imposing $\pi_C = \pi_G = \theta/2$ on the frequencies of the HKY85 model. The GG98 model consequently uses a single composition parameter for each branch and for the root, instead of three. Nevertheless, this model is still hardly tractable when a large number of species is used and the user usually has to provide a set of candidate topologies to speed up the computation.

The time heterogeneous model presented here is based on these early works. Using an independent set of composition parameters for each branch, as was done in the model N2 of Yang and Roberts (1995) and in Galtier and Gouy (1998), would certainly be more realistic but a typical RNA alignment does not contain enough variable sites to estimate the frequency parameters of a complex base-pair

[4] Another issue, mentioned by Yang and Roberts (1995), is that branch lengths do not correspond exactly to the expected number of substitutions per site anymore but this is overlooked here.

model on each branch of a phylogeny. A parameter-rich model is not necessarily a better model if there is not enough data to estimate its extra parameters accurately (Steel, 2005) and restricting the number of composition vectors seems a good and necessary trade-off to model compositional heterogeneity. Foster (2004) has shown, on specific examples, that compositional heterogeneity can be accounted for with a limited number of frequency parameters and the same approach is followed here. The composition parameters for each branch are chosen from a pool of available composition vectors. The frequency parameters in the pool and their placement on the phylogeny are both parameters of the model that are estimated during the inference process.

The approach developed here is quite similar to Foster's approach (2004) but the work is carried further. First, rooted trees are considered and the root position is not constrained to internal trifurcating nodes. New MCMC proposals were devised to cope with the presence of the root and to account for the repartition parameter that allocates the composition vectors to the branches. A priori knowledge can be used to constrain the location of the root during the inference process and this functionality is used in sections 4.5 and 4.6 to reduce the complexity of the simulations on synthetic datasets generated from known trees. It is also used in section 4.7 because the method had some difficulties to recover the root position with certainty when applied to the Tree of Life rRNA dataset analyzed in this section. Nevertheless, note that it has been suggested that there might be enough information in present-day sequences to successfully recover the location of the root with nonhomogeneous methods (Yang and Roberts, 1995; Yap and Speed, 2005), even though the exact location of the root on its branch is often poorly resolved (Galtier and Gouy, 1998).

Second, the number of composition vectors is a parameter of the model which is allowed to vary during the MCMC run. Reversible jump MCMC methods (rjMCMC) are used to add or remove frequency parameters during the inference and to determine the amount of heterogeneity evidenced by the data (Green, 1995), bypassing the need for complex model selection procedures. Another interesting advantage of using rjMCMC methods is that they account for, *i.e.*, integrate out, the uncertainty in the amount of heterogeneity while other phylogenetic parameters of interest are estimated. rjMCMC techniques have already been applied in Bayesian phylogenetics to model the variations of the evolutionary process across sites (Suchard et al., 2003), for model selection (Huelsenbeck et al., 2004), and to allow for polytomous tree topologies (Lewis et al., 2005).

Third, it is shown that using a uniform prior for the repartition of the composition vectors on the branches has some unexpected, and probably unwanted, side-effects. A hierarchical model that defines a more flexible prior on this repartition parameter is presented.

4.3 MCMC proposals

In this section, the new MCMC proposals that were implemented to cope with the particularities of the nonhomogeneous model are described. Most notably, the proposals to move between different unrooted tree topologies, introduced in chapter 3, had to be adapted to handle rooted trees. Other proposals that preserve the dimension of the parameters space are also re-introduced in this section. The "standard" MCMC algorithm and the acceptance rates described in chapter 3 are still valid and can be applied for these moves. Proposals that modify the size of the pool of frequency vectors are described later in section 4.4.

4.3.1 Proposals for continuous parameters

The proposals that were previously used to change the parameters of the substitution model, *e.g.*, frequency and exchangeability parameters, are still applicable and were used without modification here. Each frequency vector in the pool is considered independently. A uniform Dirichlet prior is put on each of them and new values are proposed by drawing from a Dirichlet distribution centered at the current values. The same prior and proposal are used for the extra frequency vector that represents the ancestral composition at the root of the tree. Note that a proposal affecting a single composition vector does not necessarily invalidate all the partial likelihoods computed in the previous iteration and that some economy can be achieved by only reevaluating the nodes between the branches concerned and the root.

It is recalled that we are not using the "rate ratios" parameterization for the exchangeability parameters in this chapter. A Dirichlet prior is put on the unique set of exchangeability parameters and a Dirichlet proposal mechanism is used to propose new values (see section 3.3). When the data are partitioned, a uniform

Dirichlet prior is used for the vector that contains the average substitution rates of each data block. Once again, new values are proposed by drawing from a Dirichlet distribution centred at the current values (see section 3.3).

Proposals to change branch lengths were not modified either but the "continuous topology change", which is triggered by proposing a negative branch length, is slightly different because the NNI proposal itself had to be modified (see below).

4.3.2 Proposals in the space of topologies

The NNI and SPR proposals were slightly modified to account for the presence of the root, which is not a trifurcating node anymore. Since these proposals have a disruptive effect on the allocation of composition parameters to the branches, it is also necessary to describe how the repartition parameter is handled when a new topology is proposed. In the following figures, red numbers are used to identify the composition vector associated with each branch. For illustrative purposes, different branches are assigned different composition parameters but these could be the same in practice.

The NNI proposal is described in Figure 4.2. An internal branch is chosen at random, let it be **R**→**E** where **R** is the parent. The second sibling branch emanating from **R** is then swapped with one of the child branches of **E** chosen randomly. The subtree containing the root (**A** in the figure) is kept unchanged. The proposal is unaffected if the chosen internal branch is directly linked to the root, *i.e.*, if **R** is the root and **A** does not exist. The Hastings ratio for such a proposal is 1.0. Note that all possible rooted topologies are theoretically accessible from any starting point using only this proposal.

The SPR proposal is described in Figure 4.3. Two branches are chosen randomly and the first branch is detached and reattached on the second one. Note that the two branches linked to the root are considered as a single branch for this proposal. When the first branch is removed, the two adjacent branches have to be merged into a single one. Their two lengths are summed but it would be difficult to design a proposal that would merge their two composition vectors while remaining reversible. Consequently, the pruned branch "carries" one of the composition vectors and inserts it on the destination branch. The Hastings ratio for such a move is not trivial and depends on the lengths of the branches involved

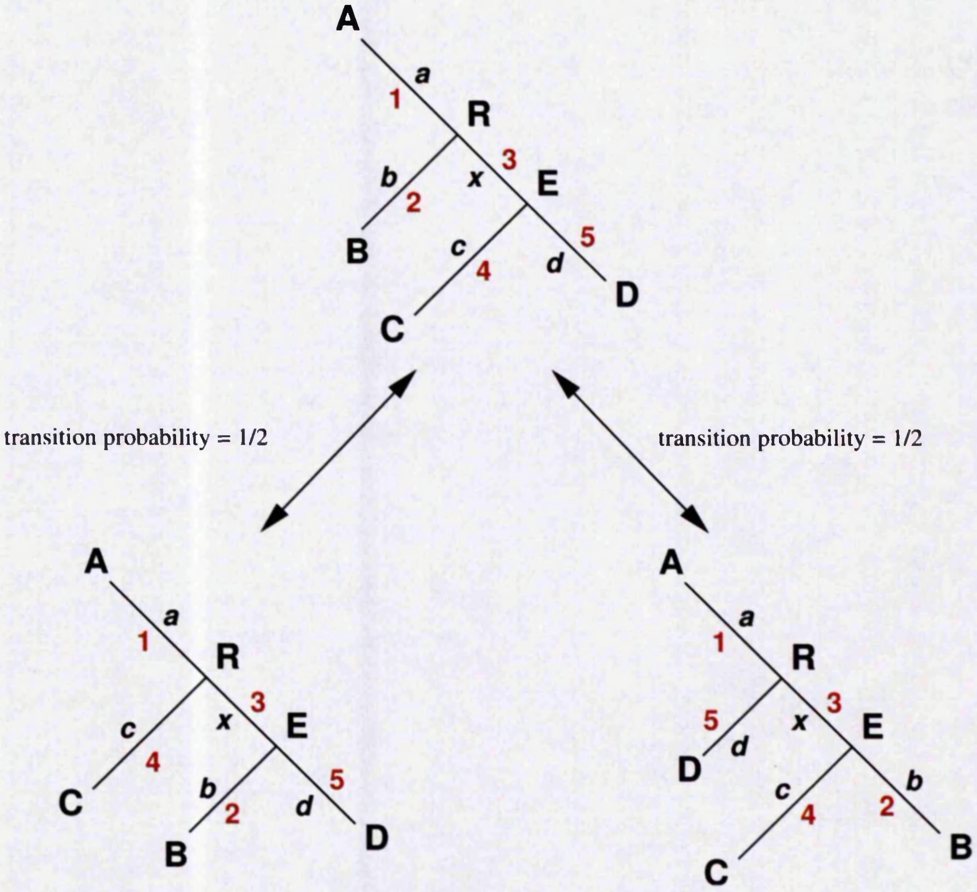


Figure 4.2: Nearest Neighbor Interchange (NNI) with the nonhomogeneous substitution model. Branch lengths are represented with lowercase letters and composition vectors with numbers.

in the move. It is the ratio of the length of the insertion branch to the sum of the lengths of the two branches adjacent to the displaced branch, which is actually the same as for the SPR proposal for unrooted trees.

4.3.3 Swapping composition vectors

This proposal changes the composition vectors assigned to a subset of branches of the tree. First, a limited number of branches is selected for the proposal. Each branch is added to the subset with probability p. Since this proposal would not modify the current state when this subset is empty, it is applied to one branch randomly chosen if no branch is selected in the first step. The composition vector of each selected branch is then replaced by another vector drawn from the pool.

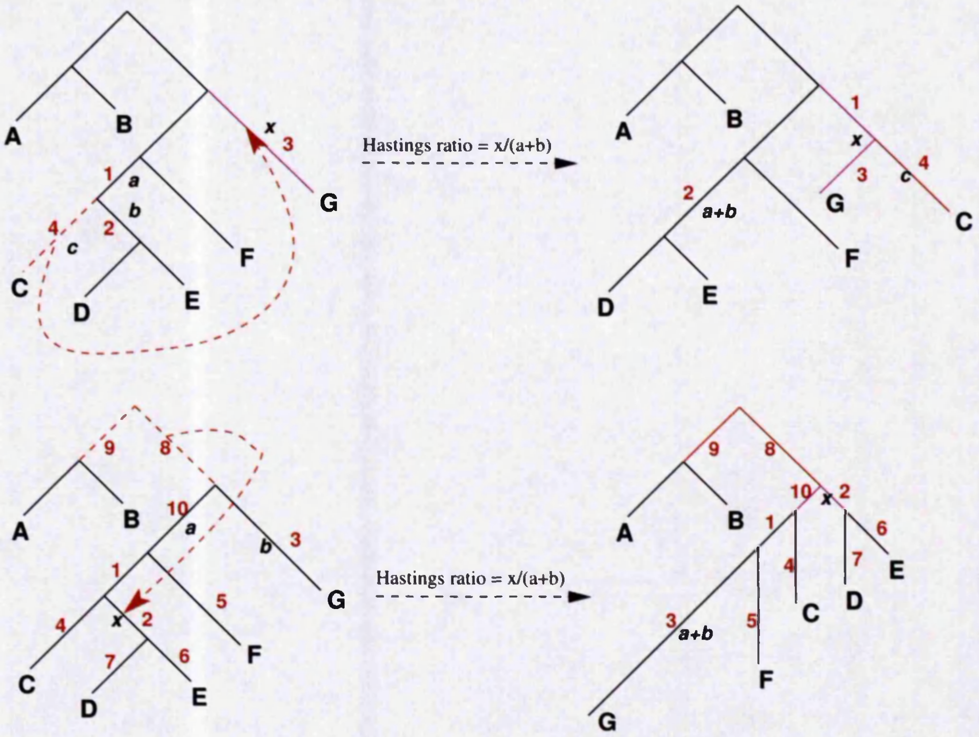


Figure 4.3: Subtree Pruning and Regrafting (SPR) with the nonhomogeneous substitution model. Branch lengths are represented with lowercase letters and composition vectors with numbers. Top)standard case: the moved subtree does not contain the root. Bottom) special case: the subtree contains the root.

Note that the current vector is excluded from the draw and this proposal is consequently not allowed when there is just one vector in the pool. p is tuned during the burnin period to reach an acceptance rate between 20% and 25%. The Hastings ratio of this proposal is 1.0. Some partial likelihoods used in the previous iteration are still valid when attempting such a proposal and partial likelihoods at an internal node do not have to be computed if the subtree below this node was not modified.

Birth and death proposals were not implemented and it is quite possible for one or more composition vectors of the pool to end up being unused with this proposal. Such superfluous parameters have no influence on the likelihood and are barely penalized by the prior with the hierarchical Bayesian model presented in section 4.5. This inflates the amount of compositional heterogeneity without real justification from the data but this is not considered as a crucial issue. The problem can easily be corrected with a prior that penalizes substitution models using a larger number of composition vectors.

4.4 Split and merge moves and the reversible jump MCMC method

4.4.1 Reversible jump MCMC computation

The MCMC techniques introduced thus far are restricted to problems where the dimensionality of the parameter space is fixed. This would not be an issue if the structure of the "true" substitution model was known with certainty but, since this is not the case, one would like the MCMC sampler to jump between the different substitution models that could reasonably explain the observed data. This would allow the inference not only to produce an estimate for the phylogeny by integrating over the possible parameter values for a specific model but also to integrate over all the possible models. Such a Markov chain would visit the different models in proportion to their posterior probability and could also be used to approximate Bayes factors used in model selection. The Bayes factor to compare model $\mathcal{M}_1$ and model $\mathcal{M}_2$ is the ratio of their evidence (see section 3.1) and can be expressed with the prior and posterior probabilities of the two models:

$$\mathrm{BF}_{1/2} = \frac{P(X|\mathcal{M}_1)}{P(X|\mathcal{M}_2)} = \frac{P(\mathcal{M}_1|X)/P(\mathcal{M}_1)}{P(\mathcal{M}_2|X)/P(\mathcal{M}_2)} \quad , \tag{4.1}$$

which is equivalent to the posterior odds when flat priors are assumed.

The MCMC principles introduced in chapter 3 cannot be used to jump between two states where the joint probability density of the parameters do not share the same underlying measure[5]. Practically, this means that a slightly different framework is needed to move between subspaces of differing dimensionality. Green (1995) described such a framework and showed how the standard Metropolis-Hastings algorithm is modified to perform reversible jumps between different spaces.

Let us assume that the models $\mathcal{M}_1$ and $\mathcal{M}_2$ define two parameter spaces

[5] Since each topology defines a different subspace for its branch lengths. NNI and SPR proposals introduced in chapter 3 are actually a special case of reversible jump proposals (Suchard et al., 2001).

of differing dimension d_1 and d_2, with $d_1 < d_2$, and that one is attempting a jump from the state $\{k=1, \theta^{(1)}\}$ to the state $\{k=2, \theta^{(2)}\}$ where k is a discrete parameter representing the current model and $\theta^{(1)}$ and $\theta^{(2)}$ are sets of parameters valid in their two subspaces. To perform a reversible jump from the first to the second parameter space and to cope with their unequal dimensions, a vector of continuous random variables $u^{(1)}$ of size m_1 is drawn from a given probability distribution and $\theta^{(2)}$ is determined from $\theta^{(1)}$ and $u^{(1)}$. Reciprocally, one jumps from the second to the first space by drawing a vector of random variables $u^{(2)}$ of size m_2 and by determining $\theta^{(1)}$ from $\theta^{(2)}$ and $u^{(2)}$. Green (1995) showed that by establishing a bijection between $(\theta^{(1)}, u^{(1)})$ and $(\theta^{(2)}, u^{(2)})$ and by matching their dimension, *i.e.*, $d_1 + m_1 = d_2 + m_2$, it is possible to perform a reversible jump between the two spaces. In practice, it is often easier not to draw any random variables when jumping from a higher dimension to a lower dimension but, for reasons that are explained below, m_2 is different from 0 in this work.

The acceptance probability for the transition (1) → (2) is:

$$P(\ \phi_{n+1}=\{k=2,\theta^{(2)}\}\ |\ \phi_n=\{k=1,\theta^{(1)}\}\ ) = \min\left(1, \frac{p(k=2,\theta^{(2)}|X)u_2(u^{(2)})}{p(k=1,\theta^{(1)}|X)u_1(u^{(1)})}\left|\frac{\partial(\theta^{(2)},u^{(2)})}{\partial(\theta^{(1)},u^{(1)})}\right|\right) \quad , \tag{4.2}$$

where u_1 and u_2 are the probability distribution used to draw $u^{(1)}$ and $u^{(2)}$. The Jacobian term arises from the change of variables from $(\theta^{(1)}, u^{(1)})$ to $(\theta^{(2)}, u^{(2)})$.

4.4.2 Dimension changing proposals for the nonhomogeneous model

In this work, the focus is on the variations of nucleotide and base-pair composition in time. State frequency parameters are the only parameters that are not tree-homogeneous but other parameters could also have been allowed to vary across branches. As implemented, the pool of parameters does not contain independent composition vectors but complete substitution models. The algorithms that would let other parameters change in different lineages (*e.g.*, exchangeability parameters, relative substitution rates between the blocks of a partition, gamma shape parameter, etc) are already implemented but, for the purpose of this thesis, all parameters but the frequencies are constrained to be homogeneous.

Although more parameters could have been allowed to vary, one important limitation of the current implementation is that the various parameters that compose a homogeneous substitution model cannot be chosen independently. Each branch is assigned a unique composite vector which contains all the parameters that are not tree-homogeneous. This limitation has some consequences for the results presented here. When sequences are partitioned before an analysis to analyze different blocks with different substitution models, each block is using independent composition parameters but they have to be considered as a single composite parameter vector when they are chosen from the pool. Practically, this means that the frequency parameters of each block are assumed to change simultaneously over time.

The split and merge proposals were designed to increase and reduce the dimensionality of the model. These proposals are described here assuming that the sequence data are not partitioned and that a unique substitution model is used at all sites, but the algorithms are easily adapted to the case of partitioned data. In what follows, k will designate the number of composition vectors available in the pool. With probability $p_S(k)$, a split is attempted and two new composition vectors, $\Pi^{(1)}$ and $\Pi^{(2)}$, are created from an initial composition vector $\Pi^{(0)}$ randomly chosen from the pool. With probability $p_M(k)$, a merge is proposed and two randomly chosen composition vectors $\Pi^{(1)}$ and $\Pi^{(2)}$ are fused to build the composition vector $\Pi^{(0)}$. In the case of partitioned data with multiple sets of composition parameters, all the frequency vectors are simply split or merged in a single proposal.

4.4.3 Split and merge moves for the phylogeny

One should aim to design proposals with a reasonable acceptance probability. Good proposals to jump between different subspaces can be designed by choosing an appropriate bijection function that accounts for the "natural" relations between the variables of the two parameter spaces. This can be achieved by drawing values close to $\Pi^{(0)}$ for $\Pi^{(1)}$ and $\Pi^{(2)}$ during a split proposal and by setting $\Pi^{(0)}$ to be a reasonable "average" of $\Pi^{(1)}$ and $\Pi^{(2)}$ during a merge move.

Since composition vectors appear and disappear during these moves, the allocation vector that assigns composition parameters to each branch has to be modified when split and merge proposals are attempted. A natural choice for a split is to reassign randomly, with equal probability, the branches previously allocated to $\Pi^{(0)}$ to $\Pi^{(1)}$ and $\Pi^{(2)}$. Symmetrically, assigning to $\Pi^{(0)}$ all the branches previously allocated to $\Pi^{(1)}$ and $\Pi^{(2)}$ seems reasonable for a merge.

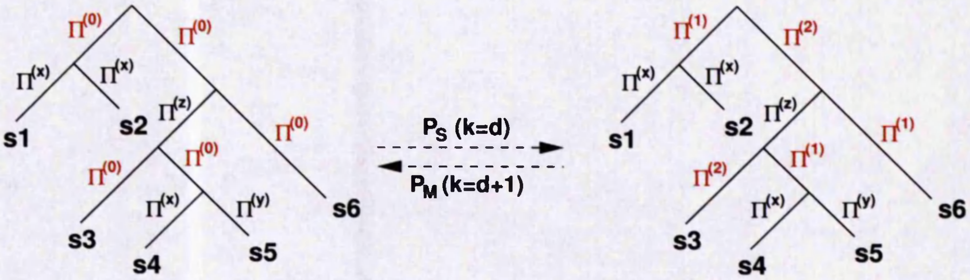


Figure 4.4: Reallocation during a split/merge proposal.

4.4.4 Split and merge moves from the pool perspective

Split and merge proposals should produce states that are similar enough to have comparable likelihood. Let b_0 be the number of branches allocated to $\Pi^{(0)}$ before the split or after the merge. Let b_1 and b_2 be the number of branches allocated respectively to $\Pi^{(1)}$ and $\Pi^{(2)}$ after the split or before the merge. Intuitively, $\Pi^{(0)}$ should be a compromise of $\Pi^{(1)}$ and $\Pi^{(2)}$ that takes b_1 and b_2 into account.

State frequencies are all greater than zero and must sum to one. The last frequency parameter is entirely determined by the others. A proposal that would match these constraints without singling out a specific frequency was designed. Dimension-matching is achieved by drawing several random variables and establishing a bijection between $\{\pi_1{}^{(0)}, \pi_2{}^{(0)}, \ldots, \pi_{n-1}{}^{(0)}, u_1, u_2, \ldots, u_n, s_0\}$ and $\{\pi_1{}^{(1)}, \pi_2{}^{(1)}, \ldots, \pi_{n-1}{}^{(1)}, \pi_1{}^{(2)}, \pi_2{}^{(2)}, \ldots, \pi_{n-1}{}^{(2)}, s_1, s_2\}$ where n is the number of frequencies in the substitution model. $\{u_1, \ldots, u_n, s_0\}$ are random variables drawn from some probability distributions when a split proposal is attempted and $\{s_1, s_2\}$ are drawn when a merge is attempted. $\pi_n{}^{(0)}$, $\pi_n{}^{(1)}$ and $\pi_n{}^{(2)}$ are determined from knowledge of the $n-1$ other frequencies.

From the three sets of frequencies $\Pi^{(0)}$, $\Pi^{(1)}$ and $\Pi^{(2)}$ and the three random variables s_0, s_1 and s_2, three sets of variables $\{m_1{}^{(0)}, \ldots m_n{}^{(0)}\}$, $\{m_1{}^{(1)}, \ldots m_n{}^{(1)}\}$ and $\{m_1{}^{(2)}, \ldots m_n{}^{(2)}\}$ are defined such that:

$$\text{for } i = 0, 1 \text{ or } 2, \quad \forall\, j \in 1..n, \qquad m_j{}^{(i)} = s_i * \pi_j \quad . \tag{4.3}$$

Consequently, $\sum_{j=1}^{n} {m_j}^{(i)} = s_i$ for $i = 0, 1$ or 2. These three sets of variables represent the unscaled composition parameters and were introduced to bypass the fact that frequency parameters have to sum up to one.

The vector $\{{m_j}^{(0)}\}$ is then defined as a compromise between the vectors $\{{m_j}^{(1)}\}$ and $\{{m_j}^{(2)}\}$ weighted by their relative importance in the phylogeny before the merge or after the split. Some models in the pool might not be allocated and there is no guarantee that b_0, b_1 and b_2 are greater than zero, which would be an issue in the mathematical expressions that follows. Therefore, $b'_1 = b_1 + 1$, $b'_2 = b_2 + 1$ and $b'_0 = b'_1 + b'_2$ are introduced to express the relations between the three vectors.

$$\forall\, j \in 1..n, \qquad b'_0 \log({m_j}^{(0)}) = b'_1 \log({m_j}^{(1)}) + b'_2 \log({m_j}^{(2)}) \quad . \tag{4.4}$$

${m_j}^{(1)}$ and ${m_j}^{(2)}$ are defined using the set of random variables $\{u_1, \ldots, u_n\}$

$$\forall\, j \in 1..n, \qquad \begin{aligned} \log({m_j}^{(1)}) &= \log({m_j}^{(0)}) + \frac{b'_0 u_j}{b'_1 \sqrt{{m_j}^{(0)}}} \quad , \\ \log({m_j}^{(2)}) &= \log({m_j}^{(0)}) - \frac{b'_0 u_j}{b'_2 \sqrt{{m_j}^{(0)}}} \quad . \end{aligned} \tag{4.5}$$

Logarithms are used to ensure that the $\{{m_j}^{(i)}\}$ values remain positive.

Admittedly, this is a rather complex choice for the bijection satisfying the dimension-matching requirement and this particular choice deserves a brief explanation. When a Dirichlet proposal is used to modify a composition vector (see equation (3.8) in chapter 3), one has to balance the desire to perform a long-distance move while maintaining a reasonable acceptance rate. The tightness parameter of the Dirichlet distribution p_0 is used to control the step-size. Experience has shown that for a fixed p_0, the acceptance rate can vary widely depending on the dataset, which is why it is modified during the burnin period to improve the mixing. Intuitively, the optimal step size for the split proposal should be related to the optimal step size used when modifying composition vectors. The bijection chosen to perform split and merge proposals was consequently designed to approximate roughly the original Dirichlet proposal. For that purpose, s_0, s_1 and s_2 are drawn from a gamma distribution with parameter p_0 and each u_j is drawn from the standard Normal distribution.

4.4.5 Computation of the acceptance rate

The acceptance probability for the split proposal is considered here.

$$A = P(\ \phi_{n+1} = \{k = d+1, \tau, \nu_\tau, {a_\tau}^{(d+1)}, \theta, \boldsymbol{\Pi}^{(d+1)}\} \mid \\ \phi_n = \{k = d, \tau, \nu_\tau, {a_\tau}^{(d)}, \theta, \boldsymbol{\Pi}^{(d)}\}\ ) \quad , \tag{4.6}$$

where k represents the size of the pool, *i.e.*, the number of composition vectors available in the current state, τ and ν_τ are respectively the current tree topology and the associated set of branch lengths as defined in chapter 2, ${a_\tau}^{(k)}$ is the allocation vector that assigns these composition parameters to the branches, θ is the constant-size vector that groups the substitution parameters that are constant across lineages, including the ancestral state distribution, and $\boldsymbol{\Pi}^{(k)}$ is the set of composition vectors available in the pool.

Reformulating equation (4.2) with a product of ratios, the acceptance rate can also be written:

$$A = \min\{1, (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio}) \times (\text{Jacobian})\} \quad ,$$

and the three last factors of this product are developed below. The likelihood is computed with the standard pruning algorithm (see section 2.6) and the first term is not detailed further here. The only important differences are that the appropriate substitution rate matrix has to be used on each branch and that the position of the root is now imposed by the evolutionary model.

The prior ratio, proposal ratio and Jacobian for the split move are developed below. Although the acceptance probability for the merge move is not explicitly given here, it can be deduced in a straightforward manner since it has the same form, with the ratio terms and Jacobian inverted, after an appropriate relabeling of k.

Prior ratio

The size of the pool, k, and the allocation vector that assigns composition parameters to the branches, ${a_\tau}^{(k)}$, are new parameters of the model. They naturally appear in the prior:

$$(\text{prior})^{(d)} = p(\tau, \nu_\tau, k{=}d, a_\tau{}^{(d)}, \theta, \mathbf{\Pi}^{(d)}) \quad .$$

A simple factorized form is chosen for this prior:

$$(\text{prior})^{(d)} = p(\mathbf{\Pi}^{(d)}|k{=}d)p(a_\tau{}^{(d)}|k{=}d,\tau)p(k{=}d)p(\nu_\tau|\tau)p(\tau)p(\theta) \quad . \tag{4.7}$$

Note that the prior on the number of substitution models $p(k=d)$ must be user-specified. A particular choice of prior on k is without consequences if Bayes factors for different values of k are the only expected results but it affects the other outputs produced by "model averaging", where k is integrated out. In this thesis, a uniform prior with an arbitrary chosen upper limit is usually used for the parameter k but a Poisson prior is sometimes used (see Figure 4.5).

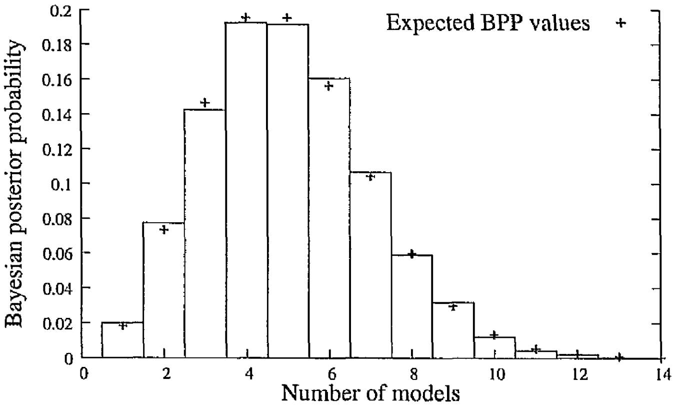


Figure 4.5: The prior on k, the number of composition vectors: algorithms described in this section were tested here with empty sequences assuming a Poisson prior distribution, $p(k',\lambda) = \frac{e^{-\lambda}\lambda^{k'}}{k'!}$, with parameter $\lambda = 4.0$ on the random variable $k' = k-1$. Without data, the inferred posterior probability of k, shown as a histogram in the picture, should be equal to the prior, shown with crosses.

For a given k, a uniform prior on the space of possible allocation vectors was chosen:

$$p(a_\tau{}^{(d)}|k{=}d,\tau) = \left(\frac{1}{d}\right)^b \quad ,$$

where b is the total number of branches in the tree. Note that this includes the configurations where some substitution models in the pool are not allocated to any branch. In the next section, it is suggested that this prior might not be appropriate and a new prior is proposed.

A simple factorized form is assumed for the prior on the parameters of the substitution model. Since most substitution parameters remain constant during a split/merge proposal, the only terms appearing in the ratio of priors are the prior probabilities of the split frequency vector $\Pi^{(0)}$, and the two resulting vectors $\Pi^{(1)}$ and $\Pi^{(2)}$. The complete prior ratio for a split move is consequently:

$$\text{(prior ratio)} = \frac{p(k=d+1)}{p(k=d)} \times \left(\frac{d}{d+1}\right)^{b} \times \frac{p(\Pi^{(1)})p(\Pi^{(2)})}{p(\Pi^{(0)})} \quad . \tag{4.8}$$

Proposal ratio

The proposal ratio is a complex term and it can be further decomposed into a product of three ratios. The first term of the product is related to the tree, more specifically to the modification of the allocation vector as described in Figure 4.4. The second term is related to the composition parameters and the probability of switching between $\Pi^{(0)}$ and $\{\Pi^{(1)}, \Pi^{(2)}\}$ once the parameters to split/merge have been selected. The third term is a non-obvious term that arises because no ordering constraint in enforced on the parameters of the different subspaces.

The probability of attempting a split and proposing $\{k=d+1, {a_\tau}^{(d+1)}\}$ from $\{k=d, {a_\tau}^{(d)}\}$ is equal to:

$$P_S(k=d) \times \frac{1}{d} \times \frac{1}{2^{b_0}} \quad ,$$

which is the probability of attempting a split, multiplied by the probability of choosing $\Pi^{(0)}$ as the composition vector to split, multiplied by the probability of choosing a specific conformation when reallocating the branches of $\Pi^{(0)}$ to $\Pi^{(1)}$ and $\Pi^{(2)}$. It is recalled that b_0 is the number of branches initially associated with $\Pi^{(0)}$ and that each branch is reallocated to $\Pi^{(1)}$ and $\Pi^{(2)}$ with equal probability $p = 1/2$. Reciprocally, from the state $\{k=d+1, {a_\tau}^{(d+1)}\}$, the probability of coming back to $\{k=d, {a_\tau}^{(d)}\}$ is equal to:

$$P_M(k=d+1) \times \frac{1}{\binom{d+1}{2}},$$

which is the probability of attempting a merge times the probability of choosing $\Pi^{(1)}$ and $\Pi^{(2)}$ as the substitution model to be merged.

The first term of the proposal ratio is consequently:

$$\frac{P_M(k{=}d{+}1) \times d \times 2^{b_0}}{P_S(k{=}d) \times \binom{d+1}{2}} \quad .$$

Let us define $P_S(k)$ and $P_M(k)$, the probability of attempting a split or a merge proposal during each MCMC cycle. Since the value of k introduces some constraints on the set of proposals that can be attempted, the expression of $P_S(k)$ and $P_M(k)$ was associated with the definition of $P_{swap}(k)$, which is the probability of the proposal that swaps the composition vector on the branches of the tree introduced in section 4.2. At each iteration, a split, merge or swap proposal is attempted with a user-defined probability q: $P_M(k) + P_S(k) + P_{swap}(k) = q$. Where it makes sense, $P_M(k) = P_S(k) = P_{swap}(k) = q/3$ can be used, but $P_S(k = 1) = q$ is compulsory since it is not possible to propose a merge nor modify the allocation vector when there is just one composition vector available in the pool. Similarly, if the prior on k defines a upper limit k_{max} on the size of the pool then $P_S(k{=}k_{max}) = 0$ and $p_M(k{=}k_{max}) = P_{swap}(k{=}k_{max}) = q/2$ can be used.

In the second term of the proposal ratio, one is concerned about the probability of switching between $\Pi^{(0)}$ and $\{\Pi^{(1)}, \Pi^{(2)}\}$. This proposal ratio is related to the probability of drawing the random variables that are necessary to perform the merge and split moves:

$$\frac{p(s_1)p(s_2)}{p(s_0)p(\{u_j\})} \quad .$$

If partitioned datasets are analyzed with different composition parameters for each block, this term appears once for each frequency vector.

The third term appears because the composition vectors in the pool are exchangeable and a particular ordering was actually assumed when composition vectors were randomly chosen before (see Cappé et al., 2003, for more details on this issue). There are $d!$ possible representations for a set of cardinality d. Consequently, the term $\frac{(d+1)!}{d!}$ appears in the acceptance probability of the split proposal to account for the ordering of the composition parameters in the pool before and after the move. Furthermore, when a split is attempted, two exchangeable composition vectors are generated and a coefficient $\frac{1}{2}$ has to appear in the acceptance probability for similar reasons.

Since the first and the third terms are almost canceling each other, the proposal ratio for the split move is simply written:

$$\text{(proposal ratio)} = \frac{P_M(k\!=\!d\!+\!1) \times 2^{b_0} \times p(s_1)p(s_2)}{P_S(k\!=\!d) \times p(s_0)p(\{u_j\})} \quad . \tag{4.9}$$

Jacobian

For a split proposal, the Jacobian of the "dimension-matching" bijection is:

$$J = \left| \frac{\partial({\pi_1}^{(1)}, \ldots, {\pi_{n-1}}^{(1)}, {\pi_1}^{(2)}, \ldots, {\pi_{n-1}}^{(2)}, s_1, s_2)}{\partial({\pi_1}^{(0)}, \ldots, {\pi_{n-1}}^{(0)}, u_1, \ldots, u_n, s_0)} \right| \quad . \tag{4.10}$$

When two or more composition vectors are used with partitioned data, the Jacobian matrix is bloc-diagonal and the Jacobian is simply the product of these independent factors. As demonstrated in Appendix A, for a single frequency vector,

$$J = \frac{{s_0}^{n-1} {b_0'}^{2n} \sqrt{\prod_{i=1}^{n} s_0 {\pi_i}^{(0)}} \exp\left(b_0' \frac{b_2' - b_1'}{b_1' b_2'} \sum_{i=1}^{n} \frac{u_i}{\sqrt{s_0 {\pi_i}^{(0)}}}\right)}{{s_1}^{n-1} {s_2}^{n-1} {b_1'}^{n} {b_2'}^{n}} \quad . \tag{4.11}$$

4.5 A hierarchical Bayesian model

4.5.1 Doubts over the uniform prior on the allocation vector

The time-heterogeneous model presented in the previous section was tested with synthetic sequences and real datasets. Results were, more or less, as expected and satisfying to an extent. Nevertheless, the posterior probability of k was generally found to be very peaked and the method seemed wary of proposing extra composition parameters unless it could impact sufficiently the likelihood value. A glance at the results suggested that the prior on the allocation vector, $p({a_\tau}^{(d)}|k = d, \tau) = (1/d)^b$, decreases very rapidly as d increases and might be a part of the issue.

Presumably, one problem is with the assumption that branches choose models independently. In practice, neighbouring branches are likely to be correlated and this results in less "entropy". It turns out that an implicit assumption of using a flat prior on the allocation vector is that composition vectors should be equally represented in the phylogenetic tree. The issue is somehow visible in the results because frequency parameters in the pool are rarely associated with a limited number of branches and tend to occupy a large proportion of the tree. When a frequency vector is allocated to a large number of branches, a slight variation of the composition parameters can have a big impact on the likelihood and can make it a worthwhile addition that balances the detrimental effect of increasing the dimension of the model on the prior on the allocation vector. However, the same slight variation on a single branch would probably go undected even though it is intuitively as relevant. This is thought to be an important issue for two reasons. First, it implies that the method is currently very sensitive to species sampling. An extra set of frequency parameters might not be added when a group is represented by a single species but the situation would probably change as more species are added. Second, compositional bias issues are, in general, invoked when a limited number of species with unusual sequence composition is found in a much larger dataset and one would want the extra composition parameters to be used to accommodate these species rather than modelling minor variations in the main part of the tree.

The problem could be fixed by modelling explicitly the changes in the model over time because changes affecting only one branch would become as likely as changes affecting large parts of the tree (see Huelsenbeck et al. (2000); Minin et al. (2005), for examples of change-point approaches in other settings). An easier solution was adopted here, by using a more flexible prior on the allocation vector. A hierarchical Bayesian model is built and the prior on ${a_\tau}^{(k)}$ is parameterized using a vector of hyperparameters $\{f_i\}$ of size k. Each f_i represents the prior proportion of branches allocated to the i^{th} composition vector and consequently $p({a_\tau}^{(d)}|k{=}d,\tau) = \prod_{l=1}^{b} f({a_\tau}^{(d)}(l))$, where $f({a_\tau}^{(d)}(l)) = f_i$ if the branch of index l is assigned the i^{th} composition vector. Elements of $\{f_i\}$ are all greater than zero and must sum to one. In this work, hyperparameters are treated like standard parameters and are estimated during the inference process. The split and merge proposals were slightly modified to account for the fact the the size of $\{f_i\}$ is not constant and to exploit at best this new parameter.

4.5.2 Changes introduced by the hierarchical model

The vector $\{f_i\}$, is handled like other frequency vectors and, keeping the dimension constant, new values are proposed using a Dirichlet distribution centred at the current value. A uniform Dirichlet prior is assumed on these hyperparameters. The prior ratio for split and merge moves, in equation (4.8), is consequently modified into:

$$(\text{prior ratio}) = \frac{p(k{=}d{+}1)}{p(k{=}d)} \left(\prod_{l=1}^{b} \frac{f^{(d+1)}(a_\tau{}^{(d+1)}(l))}{f^{(d)}(a_\tau{}^{(d)}(l))} \right) \frac{p(\Pi^{(1)})p(\Pi^{(2)})}{p(\Pi^{(0)})} \frac{p(\{f_i\}^{(d+1)})}{p(\{f_i\}^{(d)})} \quad . \tag{4.12}$$

When a split proposal is performed, one has to increase the dimension of $\{f_i\}$. In what follows, the initial proportion associated with $\Pi^{(0)}$ is designated as f_0. The two new proportions after the split, f_1 and f_2, are associated respectively with $\Pi^{(1)}$ and $\Pi^{(2)}$. An obvious choice to establish a bijection that would preserve the constraints between the set of proportions before and after the split is naturally to design a proposal that would verify:

$$f_0 = f_1 + f_2 \quad . \tag{4.13}$$

Following Richardson and Green (1997), f_0 is split into f_1 and f_2 by drawing a variable u from a *Beta* distribution $Beta(x, \gamma) = \frac{x^\gamma (1-x)^\gamma}{B(\gamma,\gamma)}$:

$$\begin{aligned} f_1 &= u \times f_0 \quad , \\ f_2 &= (1-u) \times f_0 \quad , \end{aligned} \tag{4.14}$$

where B is the beta function.

Since f_1 and f_2 are now given, it is better to use the probabilities u and $1-u$, rather than $\frac{1}{2}$ and $\frac{1}{2}$, when distributing the branches that were initially allocated to $\Pi^{(0)}$ to $\Pi^{(1)}$ and $\Pi^{(2)}$. The proposal ratio in equation (4.9) is consequently modified into:

$$(\text{proposal ratio}) = \frac{P_M(k{=}d{+}1)}{P_S(k{=}d) f_1{}^{b_1} f_2{}^{b_2}} \times \frac{B(\gamma,\gamma)}{u^{\gamma-1}(1-u)^{\gamma-1}} \quad . \tag{4.15}$$

This also introduces a supplementary Jacobian term in the acceptance rate:

$$J_f = \left| \frac{\partial(f_1, f_2)}{\partial(f_0, u)} \right| = f_0 \quad . \tag{4.16}$$

The parameter of the Beta distribution γ was chosen equal to 1.0 here.

4.5.3 Comparison of the two models

These two methods were tested with synthetic datasets that clearly reveal the differences between their priors. A random ultrametric tree of 100 species was generated with a Yule process and the common distance from root to tips was set to 0.5. On one side of the root, an outgroup clade of 12 species was found, whereas two clades $c1$ and $c2$, with respectively 18 and 70 species, can be found on the other side of the root. Using a program of the **PHASE** package, 100 alignments of 1000 nucleotides were generated by Monte-Carlo simulation using different nucleotide substitution models for the different clades. The frequency vector $\{\pi_A = 25\%, \pi_C = 20\%, \pi_G = 25\%, \pi_T = 30\%\}$ was used to generate randomly the ancestral sequence at the root and a TN93 substitution model was subsequently used to evolve the sequences along the branches of that tree (see section 2.3). A discrete gamma model with four gamma categories was assumed to model rate heterogeneity across sites and a constant gamma shape parameter equal to 0.4 was used (see section 5.2). Exchangeability parameters were tree-homogeneous ($\rho_{trans} = 0.1$, $\rho_{AG} = 0.3$, $\rho_{CT} = 0.6$), but equilibrium frequency parameters were different on the three main clades. $\{\pi_A = 25\%, \pi_C = 23\%, \pi_G = 25\%, \pi_T = 27\%\}$ was used for the equilibrium distribution on the branches of the outgroup clade, $\{\pi_A = 25\%, \pi_C = 17\%, \pi_G = 25\%, \pi_T = 33\%\}$ was used for $c1$ and $\{\pi_A = 25\%, \pi_C = 35\%, \pi_G = 25\%, \pi_T = 15\%\}$ was used for $c2$. These large alignments were then reduced to produce 100 datasets of 24 species that were analyzed with the two methods. To reduce each replicate from 100 to 24 species, the selected sequences were randomly extracted from the original dataset: 7 species were chosen from the outgroup, 5 species from $c1$ and 12 species from $c2$.

The sequences were analyzed with both methods using locally homogeneous TN93 substitution models with tree-heterogeneous frequency parameters and with a pool of variable size. A uniform prior was used for k and its upper-bound was arbitrarily set to 10. One can notice that there are comparatively few species extracted from $c1$ and that the frequency parameters used for $c1$ and the outgroup are quite similar. Both methods are consequently expected to have some difficulties to recover two different models for the outgroup and $c1$.

Results conform with the predictions. On average, the maximum a posteriori (MAP) for the number of composition vectors was found to be equal to the correct number of vectors more often with the hierarchical Bayesian model than with the original method (69% against 53%, see Figure 4.6).

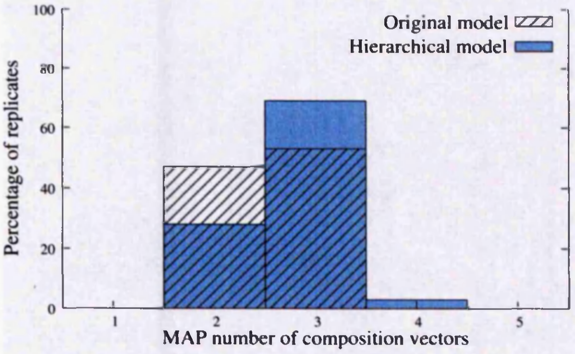


Figure 4.6: Comparison over 100 replicates of the MAP estimates for the number of composition vectors returned when using the original prior and the hierarchical prior

The posterior probability distributions for k were generally found to be peeked with the original method whereas the second method is characterized by a long tail (see Figure 4.7 for typical results with four replicates). This last issue is related to the fact that the composition vectors that are not used on the phylogeny are not strongly penalized by the prior anymore.

4.6 Bayesian simulations

The meaning of Bayesian posterior probabilities (BPPs) and their reliability as an indicator of phylogenetic uncertainty has recently attracted a lot of research effort. Bayesian support values have often been criticized on the basis that they can lead to overconfidence in incorrect nodes whereas the alternative bootstrap method, used in the ML framework, is usually more conservative and less likely to

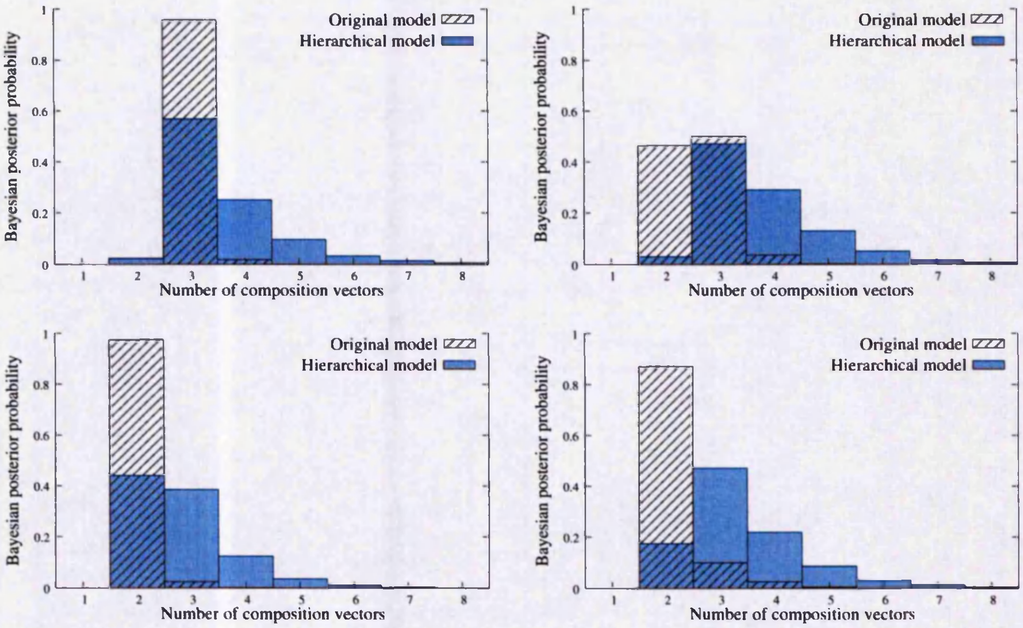


Figure 4.7: Comparison of posterior probability distributions found for the number of composition vectors by the original prior and by the hierarchical prior. Results with four selected replicates are presented here. The original model usually returns a peeked distribution whereas the hierarchical model is always long-tailed. Top-left) the original prior seems to outperform the hierarchical prior and does not introduce extra composition parameters that are not well supported by the likelihood. Top-right) the hierarchical prior seems to perform better than the original prior and clearly detects the third model. Bottom-left) both methods fails to detect the third model but the original model is over-confident on the wrong result. Bottom-right) the hierarchical prior clearly outperforms the uniform prior.

support incorrect clades (Suzuki et al., 2002; Douady et al., 2003; Erixon et al., 2003; Taylor and Piel, 2004). It cannot be denied that BPPs are usually overestimating nodal supports when the substitution model used to perform the inference is more simple than the substitution model that generated the sequences in the first place (Huelsenbeck and Rannala, 2004). Nevertheless, unlike bootstrap support values that are difficult to interpret, the BPP of a tree can readily be understood as the probability that the tree is correct, in light of the available data, if the evolutionary model is correct (Huelsenbeck et al., 2002). Some simulation studies have shown that BPPs are more accurate than bootstrap support values as a measure of uncertainty when the analysis is performed with the substitution model originally used to evolve the sequences (Erixon et al., 2003; Alfaro et al., 2003). However, these simulations have also shown that even if the generating model and the inference model are the same, BPPs for different clades do not match exactly with the actual probablity that these clades are correct for the specific evolutionary model considered during the experiments. This worrying result was explained by Huelsenbeck and Rannala (2004), who emphasized that the prior of a Bayesian analysis has to be considered as a component of the model when giving a frequentist interpretation to BPPs. Violations of the prior model can also be expected to return biased BPPs (Zwickl and Holder, 2004; Yang and Rannala, 2005).

The meaning of BPPs is not well defined when the assumptions of the model are violated. Using Bayesian simulations (Huelsenbeck and Rannala, 2004), the robustness of posterior probabilities produced by standard time-homogeneous methods is assessed when the equilibrium frequencies of the true substitution model are actually unequal in different lineages. 3000 alignments of ten species were simulated using the hierarchical time-heterogeneous model introduced in section 4.5. For each replicate, the free parameters of this evolutionary model were drawn from the prior model as defined below. The position of the root was assumed to be known in these simulations, with the first split always separating an outgroup of two species from the eight other sequences. A uniform prior was assumed on the tree topologies that match this particular constraint and each tree was randomly drawn from this prior distribution. Branch lengths were assigned by drawing values from an exponential distribution with parameter 10 (hence the expected value for each branch length is 0.1). 400 nucleotides long sequences were evolved along the branches of this tree using a time-heterogeneous TN93 substitution model. The number of composition vectors, k, was drawn from a uniform discrete prior bounded by 1 and 8 and composition vectors were subsequently drawn from a Dirichlet(2,2,2,2) distribution. A flat Dirichlet prior was used to draw $\{f_i\}$, the hyperparameters associated with the prior on the allocation vector, and these values were used to distribute randomly the different composition vectors on the branches. Ancestral frequency parameters were drawn from a Dirichlet(2,2,2,2) distribution and the unique, tree-homogeneous, set of exchangeability parameters $\{\rho_{trans}, \rho_{AG}, \rho_{CG}\}$ was drawn from a Dirichlet(5,15,30) distribution. Rate heterogeneity across sites was modelled using a gamma model with four discrete categories and the gamma shape parameter was drawn from a gamma distribution with parameters 4.0 and 0.1. The expected value of this distribution is 0.4 and its variance is 0.04.

These alignments were first analyzed with the correct evolutionary model, *i.e.*, the substitution model and the prior model used to generate the sequences. In such a case, BPPs used to measure the support of different clades are expected to match with the "frequentist" probability that these clades are correct. It is emphasized that the default priors of the **PHASE** software were not used for these simulations and were replaced by the prior model defined above. These alignments were then analyzed with the corresponding homogeneous model that uses only one homogeneous frequency vector. In both cases, the monophyletic clades found during the inference were sorted according to their posterior probability and collected into twenty bins to be compared to the frequency at which the clades in that bin were actually present in the "true" trees. Results are shown in Figure 4.8.

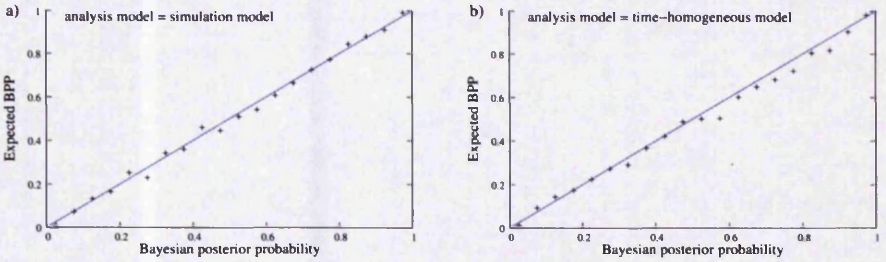


Figure 4.8: Comparison of Bayesian posterior probabilities of clades with the actual probabilities that these clades are correct. a) The model used during the analysis (likelihood model + prior model), is the model that was used to generate the data. b) The model used during the analysis is more simple and does not account for the variations of the equilibrium nucleotide frequency distribution over time.

As expected, BPPs are close to the frequentist probabilities that clades are correct when the true model is used. The results seem to suggest that the oversimplified homogeneous model has a slight tendency to overestimate Bayesian support values which is in agreement with the widespread belief that underspecified models usually overestimate clade support values (Huelsenbeck and Rannala, 2004) and supports the claim that inherited similarities in nucleotide composition can act to increase the support for correct clades, albeit for wrong reasons (Conant and Lewis, 2001). Nevertheless, this result should be considered with caution for two reasons. First, the computing power available did not allow us to perform more simulations, which would have been necessary to obtain more reliable estimates in Figure 4.8. Second, this result is only valid for a time-heterogeneous process as defined by our prior model that spreads uniformly the frequency vectors over the tree. This might not be a realistic scenario for the variation of evolutionary pressure over time and a change point model (see, *e.g.*, Huelsenbeck et al., 2000) might have returned different results.

The results also indicate that neglecting to account for the variation of equilibrium frequency over time probably has a limited impact on Bayesian support values. Nevertheless, note that the consensus trees reconstructed from the MCMC samples produced by both models were different for 871 replicates, *i.e.*, in 29% of the cases, suggesting that using a time-heterogeneous model can result in visible differences in practice. It was also noticed that the homogeneous model gave 0% BPPs to correct clades in four cases whereas the correct model was able to propose these clades with low, but non-null, supports.

4.7 Tree of Life

4.7.1 G+C content and thermophily

The G+C content of nuclear ribosomal RNA sequences is quite variable among prokaryotes and these genes would consequently constitute an interesting dataset to which one could apply the Bayesian time-heterogeneous method developed in this chapter. The small and large subunit RNA sequences of 40 species, spanning the entire biological world, were downloaded from the European ribosomal RNA database (Wuyts et al., 2004) and concatenated (see table 4.1). The alignment was manually refined to remove the highly variable regions. The secondary structures, provided individually for each sequence in the European rRNA database, were processed to produce a consensus secondary structure that defines the base-paired sites conserved in at least 50% of the sequences. The final alignment contains 3270 nucleotides separated in two blocks: 1600 unpaired nucleotides in the loops and 835 pairs in the stems.

Species	LSU acc. no	SSU acc. no
BACTERIA		
Proteobacteria		
Bartonella bacilliformis	L39095	M65249
Rhodopseudomonas palustris	X71839	L11664
Rickettsia rickettsii	U11022	L36217
Bordetella bronchiseptica	X70371	X57026
Escherichia coli	U00006	AB035925
Haemophilus influenzae	U32745	M35019
Low GC gram+		
Mycobacterium leprae	X56657	X53999
Streptomyces griseus	M76388	M76388
High GC gram+		
Bacillus subtilis	Z99104	Z99104
Lactobacillus delbrueckii	X68426	M58814
Leuconostoc mesenteroides	S60370	M23035
Staphylococcus aureus	X68425	X68417
Thermus/Deinococcus group		
Thermus thermophilus	X12612	L09659
Thermotogales		
Thermotoga maritima	M67498	M21774
Chloroplasts		
Pla. Chlamydomonas reinhardtii	X16686	J01395
Pla. Zea mays	Z00028	Z00028
Pla. Euglena gracilis	X12890	X70810
ARCHAE		
Crenarchae		
Desulfurococcus mobilis	X05480	M36474
Sulfolobus acidocaldarius	U05018	D14876
Thermofilum pendens	X14835	X14835
Thermoproteus sp	M86622	M35966

Species	LSU acc. no	SSU acc. no
Euryarchae		
Halobacterium halobium	X03407	M11583
Haloarcula marismortui	X13738	AF034619
Halococcus morrhuae	X05481	X00662
Natronobacterium magadii	X72495	X72495
Methanococcus vannielii	X02729	M36507
Methanococcus jannaschii	U67517	M59126
EUCARYA		
Metazoa		
Xenopus laevis	X59734	X02995
Homo sapiens	U13369	X03205
Fungi		
Candida albicans	L07796	X53497
Saccharomyces cerevisiae	K01048	J01353
Schizosaccharomyces pombe	Z19578	X54866
Planta		
Oryza sativa	M16845	X00755
Arabidopsis thaliana	X52320	X16077
Fragaria ananassa	X15589	X15590
Chlorella ellipsoidea	D17810	D13324
Euglenozoa		
Trypanosoma brucei	X05682	M12676
Euglena gracilis	X53361	M12677
Diplomonadida		
Giardia intestinalis	X52949	X52949
Entamoebidae		
Entamoeba histolytica	X65163	X65163

Table 4.1: Sequences and species in the TOL dataset. Sequences were downloaded from the European ribosomal RNA database (Wuyts et al., 2004) with secondary structure information.

The G+C content of nuclear RNA genes is actually correlated with the optimal growth temperature in prokaryotes (Galtier and Lobry, 1997). The rRNA sequences of thermophilic (50°C < Topt < 80°C) and hyperthermophilic species (Topt > 80°C) are G+C rich. The correlation is strongest in stem regions, presumably because **G:C** base-pairs bond together through three hydrogen bonds and are more stable than **A:U** pairs that pair with two hydrogen bonds. The additional links confer better heat resistance to the G+C rich rRNA molecules that can remain operational at higher temperature. This is illustrated in Figure 4.9 where the optimal growth temperature of the prokaryotes (Bacteria and Archaea) is contrasted with the percentage of **G:C** and **C:G** pairs found in the stems of their rRNA genes.

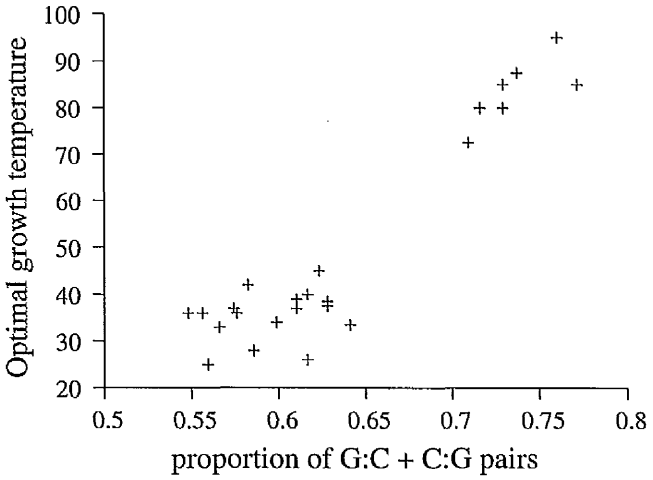


Figure 4.9: Correlation between optimal growth temperature and rRNA G+C content in prokaryotes: LSU and SSU genes from 40 species spanning the entire tree of life were aligned and concatenated. RNA stems, as defined by the consensus secondary structure that was built during the alignment process, were extracted from this alignment. Pairs of columns containing ambiguities and gaps were removed to produce a final alignment with 1226 columns (613 pairs). Using this alignment, the percentage of G:C + C:G pairs is compared with the optimal growth temperature (for prokaryotes only). OGTs can be found in Galtier and Lobry (1997) and the PGT database (Huang et al., 2004).

In section 4.1, the GG98 model that allows for varying G+C content over evolutionary time was introduced (Galtier and Gouy, 1998). By a daring use of this statistical evolutionary model, Galtier et al. (1999) argued against a hyperthermophilic last universal common ancestor (LUCA). Indeed, using a dataset similar to the dataset described above (same genes and similar species sampling), the inferred ancestral G+C composition at the root of the Tree of Life was found to be incompatible with a hot living environment. This section consequently has two aims: illustrate the time-heterogeneous method on a real example and repeat the analysis of Galtier et al. (1999) with Bayesian methods using only the stems. The original analysis returned ML estimates for the ancestral G+C composition and used complete RNA sequences, RNA loops included.

Although rRNA genes have contributed the most to the reconstruction of the Tree of Life so far, it is unlikely that using more complex evolutionary model with these genes can further resolve the deep branching relationships of the tree. Comparing the results found with time-heterogeneous methods and standard homogeneous methods can certainly bring some insights on the behaviour and differences between these two methods but it is not claimed here that the results are necessarily better when the tree topology is all that matters.

4.7.2 Phylogenetic reconstruction

The concatenated LSU+SSU dataset was analyzed with the time-heterogeneous method developed in this chapter using a combined model: TN93 for loops and 7D for stems. Frequency parameters of both substitution models were allowed to vary across branches with the restriction described above that frequency parameters for both blocks cannot be selected independently and that each branch is allocated a composite set of eleven (4 + 7) frequency parameters from the pool. In both blocks, rate heterogeneity across sites was accounted for using the discrete gamma model, see section 5.2 or Yang (1994), with eight discrete gamma categories.

Sixteen MCMC experiments were run without constraining the position of the root but the chains converged towards different results. Eleven chains converged towards a distribution where the root was on the bacterial branch with 100% BPP and three chains converged toward a distribution where the root was on the eukaryal branch with 100% BPP. The two remaining chains were discarded since the root, which was initially within the Archaea, switched to the branch leading to Bacteria during the sampling. This clearly means that current algorithms cannot resolve the position of the root on the rRNA tree and cannot attach a posterior probability to the different alternatives. Since the chains were run for 30, 000, 000 iterations, plus 9, 000, 000 iterations for the "burn-in", we do not believe longer runs could have solved the problem.

The rooting of the universal tree is not yet a resolved problem. Since the monophyly of Archaea is uncertain and not supported by this dataset, we will only consider the bacterial and the eukariotic rootings in what follows. Even though the fourteen retained chains converged towards two different posterior distributions, results were consistent for a given rooting point. These results are summarized in Figure 4.10 for the bacterial rooting and in Figure 4.11 when the root is on branch leading to the eukaryotes. These two consensus trees were produced using the extended majority rule consensus method on the set of sampled trees. Numbers in red represent Bayesian posterior probability of the corresponding clades. For a given root, these clade support values were consistent across chains. The branch lengths shown are mean posterior estimates computed using the sampled states in which the corresponding clade was present. Branches were colored with the mean G:C + C:G content of their allocated set of base-pair frequency parameters, using, once again, the samples in which the corresponding clade was present.

Apart from the root position, one can notice that both tree topologies are actually identical. This tree is similar to the tree recovered by Galtier et al. (1999) and is most probably the same (cf fig. 1 of their paper). The monophyly of Archaea is rejected in both cases, albeit with a low support for the clade Euryarchaeota-Eubacteria when the root is positioned on the branch leading to Eukaryotes. Both trees seem surprisingly well resolved, *i.e.*, not necessarily correct but with high BPPs for the different clades, most of them being equal to 100%. Nevertheless, clade posterior probabilities were found to be even higher when the corresponding homogeneous model, *i.e.*, TN93+7D with uniform frequency parameters, was used (tree not shown) and these results are actually not so surprising. The unrooted topology recovered with a time-homogeneous model was found to be the same except for one major difference: when frequency parameters are assumed to be constant across lineages, *Entamoeba histolytica* is found to branch out after the phylum Euglenozoa (represented here by two species). This suggests that the branch leading to the G+C rich *Giardia intestinalis* and the branch leading to the G+C poor *Entamoeba histolytica* are repulsing each other when variation of the substitution process over time is not accounted for (see also Hasegawa et al., 1993).

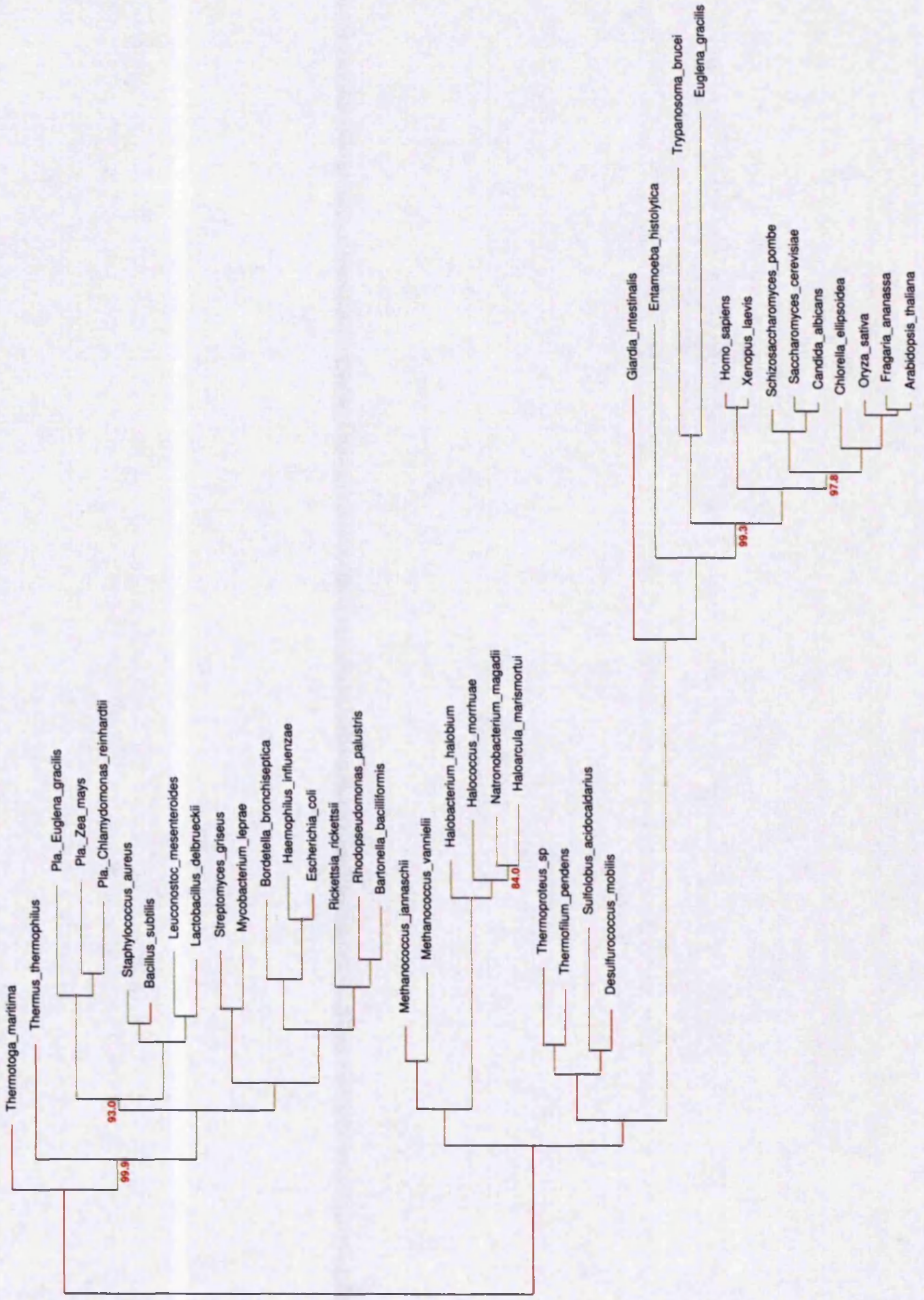


Figure 4.10: Consensus tree produced from the sampled states of the 11 chains that converged towards a bacterial rooting. Numbers in red are the BPPs for the corresponding clades, no number means 100%. Colors represent the G:C+C:G equilibrium frequencies of the base-pair substitution model, ranging from 13% (light green) to 86% (bright red). Branch lengths and equilibrium frequencies were averaged over the samples for which the underlying clade was present.

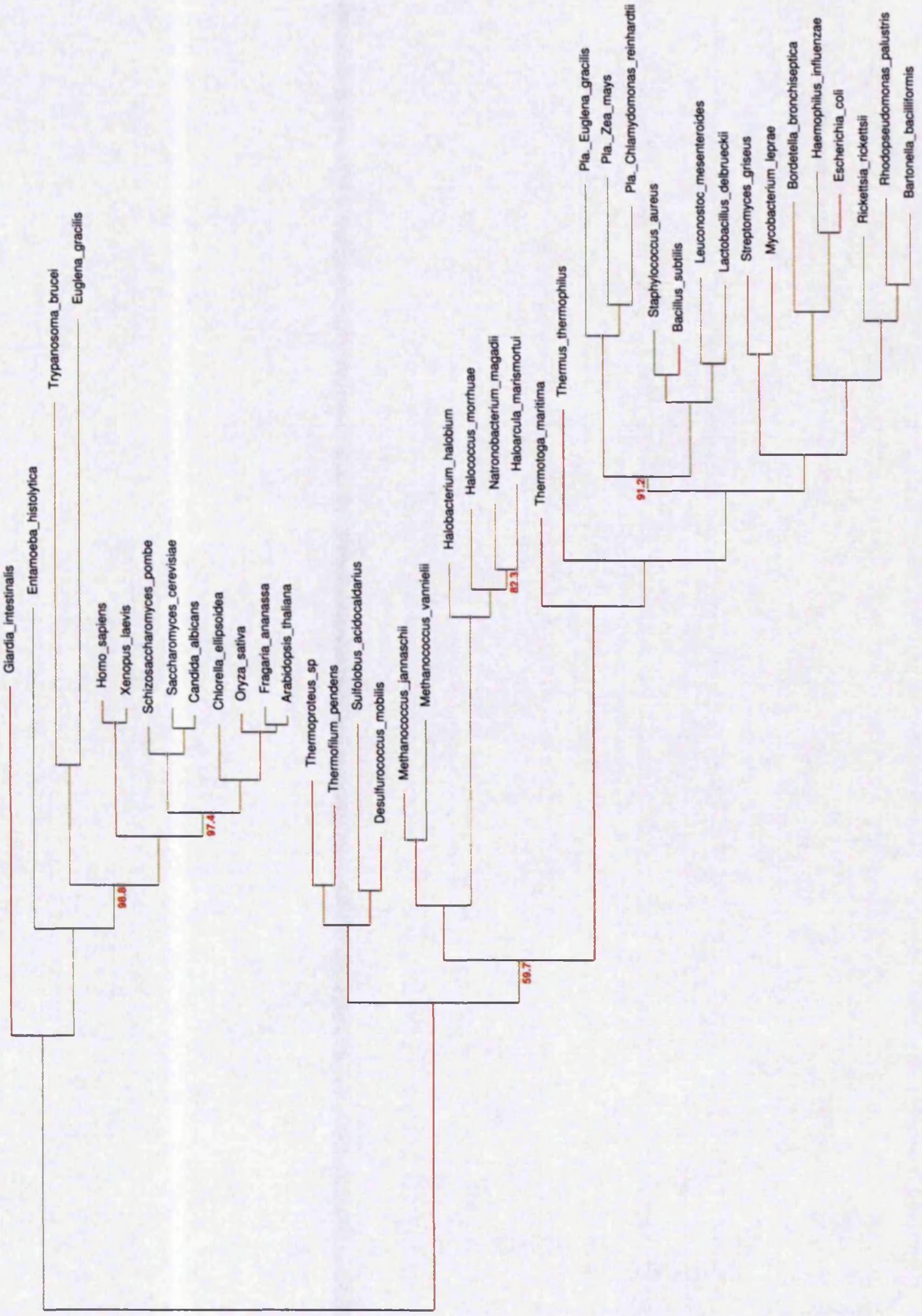


Figure 4.11: Consensus tree produced from the sampled states of the 3 chains that converged towards an eukaryotic rooting. Numbers in red are the BPPs for the corresponding clades, no number means 100%. Colors represent the G:C+C:G equilibrium frequencies of the base-pair substitution model, ranging from 13% (light green) to 86% (bright red). Branch lengths and equilibrium frequencies were averaged over the samples for which the underlying clade was present.

4.7.3 Ancestral G+C content

Using their software **EVAL_NH**, Galtier et al. (1999) recovered a ML estimate for the ancestral G+C content of LUCA that falls at the upper end of the range of G+C content found in contemporary mesophiles (cf fig. 2 of their paper). They consequently concluded that LUCA was probably not a hyperthermophile. This result was found to be quite robust to the gene used (*i.e.*, SSU or LSU), to the rooting point and to species sampling. We partially repeated their experiment using the Bayesian method developed in this chapter. Since the variation of the G+C content is related to the variation of the proportion of G:C pairs in RNA stems and is only weakly correlated to the G+C content in RNA loops (Wang and Hickey, 2002), loops were not used during this analysis. To allow for direct comparison between the inferred ancestral G:C+C:G content and the G:C+C:G content observed in contemporary species, positions containing pairs with gaps and/or ambiguous nucleotides were also removed from the dataset, leaving 613 aligned pairs.

We inferred the ancestral G:C+C:G content assuming that the unrooted topology found with the complete dataset was correct (see Figure 4.10 and Figure 4.11). The position of the root was constrained during these analyses but both rooting points were considered. A 7D substitution process with eight discrete gamma rate category was used to model the evolution of RNA pairs and four MCMC runs were performed for each possible rooting point. Since this dataset is smaller, we could afford $60,000,000$ sampling iterations for each run. The inferred ancestral G:C+C:G contents are given in Figure 4.12 and compared with the G:C+C:G contents of extant species.

With the bacterial rooting, the mean posterior estimate for the ancestral G:C+C:G content is $\text{G:C}_{bact} = 66.02\%$. With the eukariotic rooting, $\text{G:C}_{euk} = 61.42\%$ was inferred. While these results do not drastically differ from the results of Galtier et al., one can notice that the Bayesian credibility intervals recovered here are not as tight as the confidence intervals proposed in Galtier et al. (1999). This increase in variance is not totally unexpected since loops were not used and the dataset is smaller. More importantly, results with the eukaryotic and bacterial rooting are found to be quite different in these analyses (see Figure 4.13). A mesophilic LUCA is not unequivocally supported when the bacterial rooting is used and most of the posterior distribution is falling in an area without data

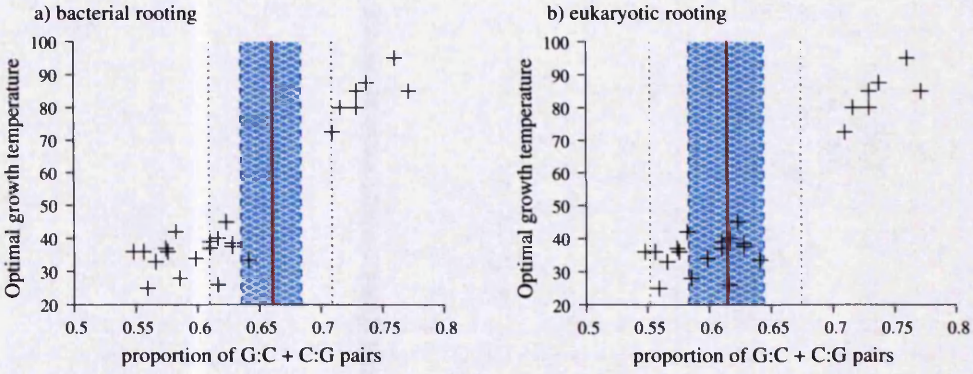


Figure 4.12: Ancestral G:C content inferred with the Bayesian time-heterogeneous model applied on concatenated rRNA stems assuming a) a bacterial rooting, b) an eukaryotic rooting. Red lines represent mean posterior estimates, boxes represent 66% credibility intervals and dotted lines are the limit of the 95% credibility intervals.

points. The observed difference between the two rootings is consistent with the fact that species at the base of the bacterial clade are thermophilic.

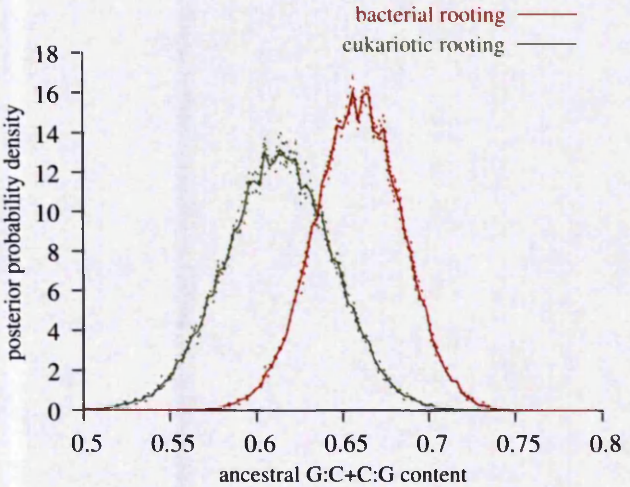


Figure 4.13: Posterior probability densities for the ancestral G:C contents (bacterial and eukariotic rootings) previously shown in Figure 4.12. Results for each MCMC run are shown with dotted lines to indicate that the independent chains converged to a stable and consistent probability distribution which is different for the two possible roots.

Experiments were repeated without fixing the topology in order to account for phylogenetic uncertainty when inferring the ancestral G:C+C:G contents. The bacterial and eukariotic clades were assumed to be monophyletic and the ancestral

G:C+C:G content was estimated once again using both possible rootings. The two consensus trees, inferred only with the stems, are slightly different and the support values are lower than with the complete dataset. Nevertheless, mean posterior estimates and credibility intervals for the ancestral G:C+C:G content are found to be almost identical (G:C_{bact} = 65.97% and G:C_{euk} = 61.45%) and accounting for the phylogenetic uncertainty does not alter the previous results.

These analyses differ from the original experiments of Galtier et al. (1999) in many points. Nevertheless, there is no obvious reason that could explain why using a base-pair substitution model with only the stems returns different results. This problem is approached again in the next chapter and a possible explanation for these divergences will be proposed.

4.7.4 Amount of heterogeneity

The number of composition vectors is an important parameter of the time-heterogeneous substitution model developed here. Previous models developed in the ML framework had many free composition vectors, which prevented their use with a large number of species unless the user could restrict the number of candidate phylogenies before the analysis. Based on the notion that compositional variation may be a rare event and that equilibrium frequency parameters might be homogeneous on large parts of the tree, Foster (2004) proposed a tractable model that could potentially account for compositional heterogeneity with a limited number of parameters if permitted by the (lack of) heterogeneity evidenced in the data.

When analyzing the LSU sequences of 5 bacterial species (see Figure 4.1), we found that only 2 composition vectors could explain the data well, which is consistent with Foster's result using the SSU gene. Nevertheless, once a flexible prior was implemented for the repartition of composition parameters on the tree (see section 4.5) and with a larger number of species, it was found that a relatively large number of composition vectors were necessary to fit the Tree of Life dataset presented here (see Figure 4.14). This suggests that the method cannot be applied easily to large dataset when the data are highly heterogeneous. The chains had to run for a long time to produce the results presented here.

As hinted at the beginning of this chapter, the choice of a particular prior on the number of frequency parameters can have a substantial impact on the

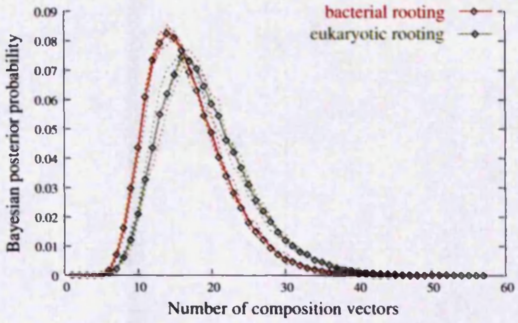


Figure 4.14: Posterior probability distribution for the number of composition vectors k, in the time-heterogeneous 7D model used to analyze the stems and infer the ancestral G:C+C:G content in the previous section. Results for each independent MCMC run are shown with dotted lines. The arbitrary choice of a particular rooting has a visible effect on the results. We recall that a uniform prior with an arbitrary upper-bound is used by default for k.

posterior distribution. In Figure 4.15, the posterior distributions on the number of composition vectors obtained with various initial prior distributions are compared. Using a Poisson prior with a low mean significantly affects the posterior distribution on the number of frequency vectors but the impact on the ancestral G:C+C:G posterior distribution is once again not detectable. Clade support values were also largely unaffected by a change of the prior distribution (differences were within 6%).

4.8 Conclusion

Current ML methods that allow for compositional variation over evolutionary time do not scale well when a large number of species is used because they are using independent composition vectors on each branch. Large trees have many parameters and this complicates tremendously the search for the ML topology. Following Foster's approach (2004), a time-heterogeneous model using a limited number of frequency parameters was proposed in this chapter. The method can infer phylogenetic trees with nonhomogeneous substitution models and remains computationally tractable. An important aspect of this work is that the amount of heterogeneity does not have to be *a priori* specified by the user or determined with complex model selection procedures from the results of multiple chains.

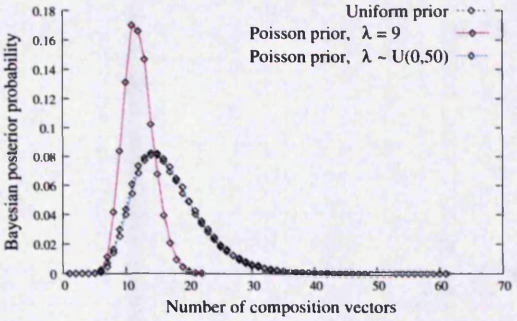


Figure 4.15: Posterior probability distribution for the number of composition vectors in the time-heterogeneous 7D model used to analyze the stems and infer the ancestral G:C+C:G content in the previous section. Results are presented for the bacterial rooting. The dotted line represents the results with a uniform prior previously shown in Figure 4.14. The two other curves were found when assuming a Poisson prior distribution ($p(k', \lambda) = \frac{e^{-\lambda}\lambda^{k'}}{k'!}$) on $k' = k - 1$. In the first case, λ was fixed to 9 and in the second case λ was a hyperparameter of the model and was given a uniform prior distribution.

The act of adding and removing frequency parameters is handled during the inference process using reversible jump methods and the posterior probability for the number of compositon vectors can easily be computed from the outputs of the MCMC sampler. More importantly, the uncertainty in the amount of heterogeneity is integrated out when estimating other phylogenetic parameters of interest. It was also shown in this chapter that using a flat prior on the allocation vector that distributes the frequency parameters over the branches of the tree may lead to unexpected results and will probably cause the method to underestimate the amount of heterogeneity evidenced by the data. Finally, when the method was used to replicate Galtier et al.'s (1999) results using RNA stems only, it was not possible to completely rule out a thermophilic ancestor, although an hyperthermophilic one seems unlikely.

The methods that were previously developed to account for the variation in nucleotide frequencies across lineages do not always solve the obvious topological errors encountered when compositionally heterogeneous datasets are studied (Foster and Hickey, 1997; Chang and Campbell, 2000; Tarrío et al., 2001). Consequently, the method developed here, which is in essence very similar, will probably not be the panacea for all composition-related reconstruction artifacts. Conant and Lewis (2001) have already argued that compositional heterogeneity alone could not completely explain the failure of traditional methods in some examples used by Lockhart et al. (1994). This suggests that accounting only for the compositional bias cannot prevent all reconstruction artifacts. Frequency parameters are probably not the only one changing over evolutionary time. Alternative methods that circumvent the bias by recoding the data — *e.g.*, RY-coding (Woese et al., 1991; Phillips et al., 2004) or AGY-coding (Gibson et al., 2005) — might be a better strategy with some datasets.

Chapter 5

Heterogeneity across sites

Model-based phylogenetic reconstruction methods traditionally assume homogeneity of nucleotide frequencies among lineages but they also assume homogeneity across sequence sites. Compositional variation in time has already been extensively studied but few studies have focused on the effects of compositional heterogeneity across sites. It is demonstrated here that different sites in an alignment do not always share a unique compositional pattern. Specific examples where compositional trends are correlated with the site-specific rate of evolution in RNA genes are provided. Compositional heterogeneity across sites is shown to perturb the estimation of evolutionary parameters with standard phylogenetic methods and also affects the ancestral composition estimate returned by time-heterogeneous methods. The latter finding could challenge the results of Galtier et al.'s study (1999) arguing against a hyperthermophilic last universal ancestor and could explain the slightly contradictory results found in the previous chapter. A new model is proposed to account for compositional variation across sites. A Gaussian process prior was designed and used to allow for a smooth change in composition with evolutionary rate in the Bayesian framework. The results suggest that this model can accurately capture the observed trends in present-day RNA sequences.

5.1 Introduction

Early phylogenetic methods assumed that all sites in a molecular sequence evolve according to the same pattern. Since the selection pressure does not act uniformly over the length of a gene, this assumption is likely to be strongly violated with real data. In an attempt to capture evolutionary information more accurately, researchers incorporated more complexity, and biological reality, in their evolutionary models. The introduction of among-site rate variation (ASRV) in substitution models has proved to be an important step towards more realistic evolutionary models (Yang, 1996a; Felsenstein, 2001). Indeed, the rate at which a mutation is fixed, or filtered out, in a population clearly depends on its position in the sequence. Some sites change often over evolutionary time whereas others are almost invariant due to strong functional constraints (Kimura and Ohta, 1974).

There are a number of approaches that can be used to accommodate for ASRV. Nevertheless, as seen in section 5.2, most of these methods only account for the variation of the speed of the nucleotide replacement process whereas other evolutionary patterns are still assumed to be shared across sites. In other words, the main parameters of the Markov process (namely equilibrium frequencies and exchangeabilities) are assumed to be constant for the whole alignment and traditional ASRV models simply act on the multiplying factor μ used to scale the transition matrix Q by letting it vary across sites (see equation (2.10)). An important point developed in this chapter is that the combined effects of mutation and selection also have an impact on nucleotide frequencies that cannot be captured by current ASRV models.

There are cases where evolutionary patterns are known to be different within a gene and where variations can easily be associated with specific DNA regions (*e.g.*, different codon positions of a protein coding gene or loops and stems of RNA genes). As was done in previous chapters, the easiest way to account for this heterogeneity is to partition the data beforehand according to *a priori* knowledge and to use combined models that use different substitution processes for different partitions. Nevertheless, partitioning of the data is only an option when the different blocks can be identified in advance and the problem tackled by this chapter is actually the detection and accommodation of compositional heterogeneity within each block of a partition.

In the past ten years, methods have been proposed to capture spatial heterogeneity (and not specifically compositional heterogeneity) when the correct partitioning scheme is unknown. Latent class models have been used quite often in this context. These models assume that each site is evolving according to an unobserved process chosen among a finite set of substitution models and, as briefly mentioned in section 2.6, the likelihood is computed by integrating over all the possible substitution processes at a site (see equation (2.18)). For instance, Huelsenbeck and Nielsen (1999) introduced a model where the transition/transversion rate ratios can vary along a sequence according to a gamma distribution. Following Nielsen and Yang's (1998) work, Yang et al. (2000) proposed a set of codon models with variable synonymous/nonsynonymous ratio across sites. The standard discrete gamma rate model for ASRV (Yang, 1994) and the invariant sites model (Reeves, 1992), used in previous chapters and introduced in section 5.2 below, are early examples of latent class models that allow for rate-heterogeneity across sites without prior classification.

Closer to the main topic of this chapter, some methods have also been devised specifically to account for variation in composition across sites. Most substitution models ignore this aspect of sequence evolution and assume that the equilibrium frequencies across sites are constant. Dimmic et al. (2000) incorporated site-specific selection effects using different amino-acid fitness functions. More recently, Pagel and Meade (2004) introduced a "pattern-heterogeneity" mixture model to account for the heterogeneity both in average evolutionary rate and exchangeability parameters, and their model can easily be extended to describe the spatial variation of frequency parameters. Both models are also latent class methods. Models using a specific frequency vector at each site have also been designed. Lartillot and Philippe (2004) constructed a model that allows for a variable number of frequency profiles to fit the variation of the amino-acid equilibrium distribution and, in effect, classifies sites into distinct categories. A parameter-rich model was also attempted by Bruno (1996) and Halpern and Bruno (1998), where a single mutation process is shared across sites but a site-specific frequency vector is used to model the variation of selection effects across sites.

In this chapter, the focus is on the variation in nucleotide composition across RNA genes. Once again, secondary structure constraints and differences between paired and unpaired regions are taken into account. By contrasting the composition observed at slow and fast evolving pairs in RNA stems in section 5.3, it is suggested that the differences between their patterns of evolution is not simply limited to variation of the average substitution rate. The most stable G:C pairs are more common in slowly evolving parts of RNA stems and nucleotide frequencies in RNA loops are also found to be quite variable across rate categories. In general it appears that base composition is correlated with the site-specific evolutionary rate.

To motivate the work, the biasing effects of compositional heterogeneity across sites are highlighted in section 5.4. In previous works, it has been noticed that frequency estimates were biased towards the frequencies of fast evolving sites (Jow et al., 2002; Hudelot et al., 2003) and it is investigated whether compositional variations across sites could explain these unexpected results. More generally, the behaviour of standard phylogenetic methods in presence of compositional variation across sites and the detrimental impact it has on various phylogenetic estimates are explored.

In order to account for the observed compositional trends across sites in present-day sequences and avoid the forementioned issues, a new latent class method for ASRV is introduced in section 5.5. The proposed method extends the discrete gamma model (Yang, 1994) to allow for variable equilibrium frequencies in each rate category. The main model assumption, supported by empirical evidence, is that these frequencies change in a smooth way across rate categories. Gaussian processes are used to control the smoothness and parameters are estimated from the data in the Bayesian framework using MCMC techniques. This model is implemented and available in **PHASE**.

As documented in chapter 4, more effort has been put into modelling compositional variation over evolutionary time rather than across sites (Yang and Roberts, 1995; Galtier and Gouy, 1998; Brooks et al., 2004; Foster, 2004), presumably because lineage-specific base compositional bias has been shown to cause worrisome phylogenetic artifacts. Nevertheless, neglecting to account for compositional variation across sites can also have worrying consequences on models proposed to relax the assumption of time-homogeneity, including the model introduced in the previous chapter. As an example, it is shown in section 5.6 that under a model that wrongly assumes across-site homogeneity of the equilibrium frequencies, part of the observed variation across time will in fact be due to variations across sites. Galtier et al. (1999) suggested that the low inferred G+C content of the Last Universal Common Ancestor (LUCA) of all extant life forms was not compatible with the expected G+C content of a thermophilic species but this result has been challenged by several other studies (Di Giulio, 2000, 2003; Schwartzman and Lineweaver, 2004). Results presented here suggest that the ancestral G+C compositions proposed by Galtier et al. (1999) for the two studied rRNA genes were most likely underestimated.

5.2 Modeling rate variation across sites

A simple method to account for ASRV is to assume that there exists a fixed number of rates at which a site can possibly evolve. When the assignments of sites to the different rate categories is unknown, the marginal likelihood at site j is computed using the weighted sum over all possible assignments as in equation (2.18),

$$\begin{aligned} L_j &= P(X_j|\tau, \nu, \theta) \\ &= \sum_{c=1}^{C} P(X_j|\tau, \nu, \theta, r_c) p(r_c) \quad , \end{aligned}$$

where C is the number of rate categories, $p(r_c)$ is the probability that a site evolves at rate r_c and $P(X_j|\tau, \nu, \theta, r_c)$ is the likelihood calculation at site j assuming that it evolves at rate r_c.

Even though the method is straightforward, there still remains the problem of choosing the appropriate number of rate categories and, more importantly, their respective rate. The issue is resolved by using parameterized models of ASRV. Their extra free parameters can also be determined during the inference process.

5.2.1 Invariant sites model

Perhaps the simplest way to account for rate variation across sites is to have two categories of sites. The first category contains the sites that are invariant whereas the other has a non-zero rate of evolution. A single parameter is added and controls the proportion of invariant sites $p(r_c = 0)$. The proportion of sites in the second category is naturally $1 - p(r_c = 0)$. Recall from section 2.5 that the average substitution rate $E(r)$ for a set of sites is either fixed to 1.0 or is a parameter of the model when data are partitioned. Consequently, the average rate for the second category is determined by this constraint and is equal to: $E(r)/(1-p(r_c=0))$.

5.2.2 Discrete gamma model

Yang (1993) did not use a finite number of categories but modelled ASRV by assuming that the evolutionary rate at a site was drawn from a continuous gamma distribution:

$$p(r|\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}e^{-\beta r}r^{\alpha-1} \quad , \tag{5.1}$$

with mean $E(r)=\alpha/\beta$ and variance $V(r)=\alpha/\beta^2$. Since the mean is already determined by other constraints, the gamma shape parameter α is the only parameter that controls the variance of this distribution. A small α value indicates rather large differences between sites with few sites having high rates of evolution and most sites being practically invariant (see Figure 5.1). The gamma model collapses to the single rate model with $\alpha=+\infty$.

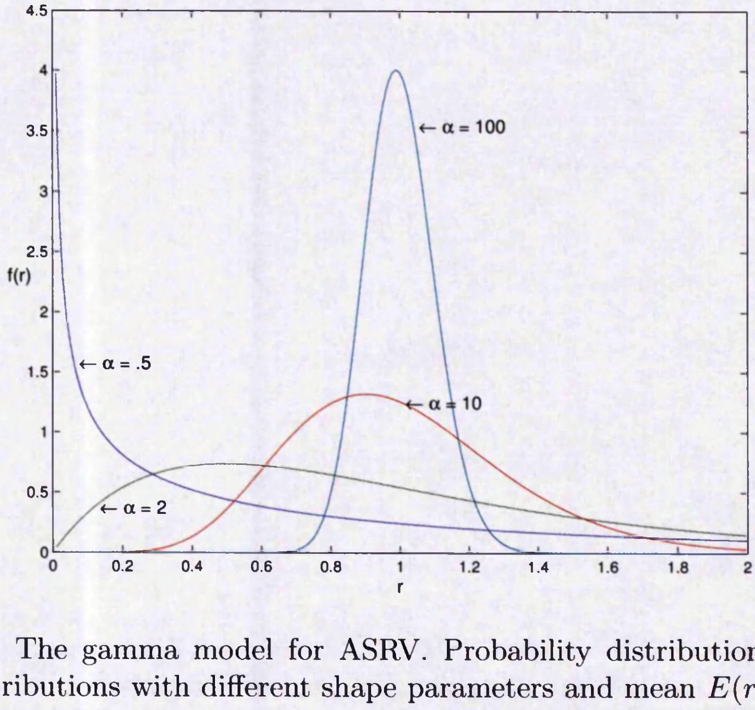


Figure 5.1: The gamma model for ASRV. Probability distributions of several gamma distributions with different shape parameters and mean $E(r)=1$

The marginal likelihood at a site can be computed by integrating over all possible rates:

$$L_j=\int_0^{\infty}P(X_j|\tau,\nu,\theta,r)p(r|\alpha)dr \quad ,$$

but it involves heavy computation. For this reason, Yang (1994) proposed an approximate method and the continuous rate distribution is substituted by a discrete one. The gamma distribution is split into C equal portions with $C-1$ cutting points and the mean rate for each portion is then used as the rate for each discrete gamma category. The rate at a site is then described as a random draw with equal probability of being in each category ($\forall\ c,\ p(r_c) = 1/C$). This gives the discrete gamma model that is used in this thesis. The discrete gamma model can be combined with the invariant sites model described above by assuming that the variable sites evolve according to a gamma distribution with mean $E'(r) = E(r)/(1-p(r_c = 0))$. This gives the popular +dG+I model which is also implemented in the **PHASE** software.

5.2.3 Allowing for auto-correlation between adjacent sites

Felsenstein and Churchill (1996) generalized the approach of Yang (1994) by introducing a dependence between the rates at adjacent sites. A hidden Markov model (HMM) is used to model the auto-correlation of sites and the rate category at site $j+1$ is supposed to be randomly drawn from a distribution that depends on the rate category previously drawn for site j. This is biologically justified by the fact that functional constraints usually act on regions containing more than one nucleotide and that contiguous nucleotides are usually under similar functional constraints.

This approach is further generalized by Thorne et al. (1996) and in subsequent works from the Goldman group (*e.g.*, Goldman et al. (1998); Liò and Goldman (2002)) where a complex HMM modelling the different structures found in protein sequences (*e.g.*, alpha-helix and beta-sheet) is trained and used to model accurately the variation of the substitution patterns across sites and is not limited to ASRV. Models of this type were considered for RNA molecules but it was not clear how to properly handle the fact that the two nucleotides constituting a base-pair are apart from each other in the sequence.

5.3 Variation of equilibrium frequencies across

sites: empirical evidence

The substitution models used in **PHASE** and in other phylogenetic software are usually parameterized with the equilibrium state distribution but the treatment of these parameters is software-dependent. Since stationarity is assumed, a sensible method, applied by default in most ML phylogenetic programs, is to obtain the frequency parameters directly from the empirical composition of the studied sequences (Whelan and Goldman, 1999). Since the cost associated with the estimation of additional parameters is not as high with MCMC techniques, Bayesian inference programs, **PHASE** included, usually consider frequency parameters as parameters to be estimated during the inference process.

In Jow et al. (2002) and Hudelot et al. (2003), it was noticed that the frequency parameters estimated during the inference process were markedly different from the empirical composition, suggesting that one or more assumptions of the evolutionary model were not met. In both works, a seven-state base-pair model was used with the stems of mammalian mitochondrial RNA genes and the estimated equilibrium frequency for the MM state was found to be much higher than the empirical proportion of mismatches (approximately 10% vs 4%). It was argued that these differences were side-effects of using an ASRV model when slow and fast evolving sites show different state distribution and this claim is investigated in the two next sections. Using a subset of the original dataset used in Jow et al. (2002) and Hudelot et al. (2003), it is shown here that the empirical composition observed at slow and fast evolving sites can indeed be different in real datasets. In section 5.4, it is shown that compositional variation across sites is responsible for the observed differences between empirical and estimated base frequencies.

5.3.1 Estimating the site-specific rate of evolution and the composition for each rate category

To estimate the composition at fast and slow evolving sites, it is necessary to categorize them in the first place. For that purpose, a likelihood-based method is designed and used to estimate the site-specific average rate of evolution and to investigate possible associations between evolutionary rates and site-specific base composition in a set of aligned sequences. The possibility of significant rate change at a site over evolutionary time is neglected here (the covarion-hypothesis introduced by Fitch and Markowitz (1970)).

It is assumed that the sequences evolved along the branches of a known tree topology (τ) with a stationary time-homogeneous model. Substitution patterns are supposed to be constant across sites but ASRV is accounted for using a discrete gamma distribution of rates. Using the given tree topology, one can compute ML estimates (MLEs) for the branch lengths ($\hat{\nu}$) and the free parameters of the substitution model ($\hat{\theta}$). These values maximize the probability that the assumed evolutionary model generated the observed sequences $P(X|\tau, \theta, \nu)$.

In an empirical Bayesian approach (Carlin and Louis, 2000; Yang and Wang, 1995), these MLEs can then be used to compute the posterior probabilities of each site j being in a specific rate category c,

$$P(r_c|X_j, \tau, \hat{\theta}, \hat{\nu}) = \frac{P(X_j|r_c, \tau, \hat{\theta}, \hat{\nu}) \times P(r_c)}{\sum_{c=1}^{C} P(X_j|r_c, \tau, \hat{\theta}, \hat{\nu}) \times P(r_c)} \quad , \tag{5.2}$$

where C is the number of rate categories and X_j the data observed at column j of the alignment. As mentioned before, rate categories have equal prior probability when the discrete gamma model is used and $P(r_c) = 1/C$.

To evaluate the equilibrium composition of each rate category, the distribution of frequencies conditional on rate is estimated using the posterior probabilities from equation 5.2 as weights. Then the expectation value of the frequency vector Π conditional on rate r_c is,

$$E(\Pi|r_c) = \frac{\sum_j E(\Pi|X_j) P(r_c|X_j)}{\sum_j P(r_c|X_j)} \quad , \tag{5.3}$$

where $E(\Pi|X_j)$ is approximated by the observed frequencies at each site.

The method supposes that the phylogeny that generated the sequences is known. Although it might have been better to take phylogenetic uncertainty into account when estimating the site-specific evolutionary rate, it is very unlikely that the arbitrary choice of a particular tree has a significant impact on the results (Mayrose et al., 2005). Substitution parameter estimates are generally found to be quite robust to the assumed topology.

5.3.2 Robustness of the method and results on artificial data

There is somewhat of a contradiction in estimating the variation of compositional patterns across sites by using a method which assumes that nucleotide frequencies are uniform along the length of the sequences. In order to ascertain whether this method is robust, it was tested with simulated datasets. Two evolutionary models, *S_homo* and *S_hetero*, were used to generate the artificial alignments used throughout this chapter and they are now described.

With *S_homo*, a standard time-homogeneous TN93 substitution model is used to generate the sequences (see section 2.3). *S_homo* is not strictly homogeneous across sites because ASRV is simulated with the discrete gamma model. Nevertheless, *S_homo* can be considered "pattern-homogeneous" because frequency parameters are constant across sites and the base frequency distribution is expected to remain homogeneous both across sites and lineages. *S_hetero* is similar to *S_homo* in all respects but, by contrast, a different set of frequency parameters are used for each gamma category. *S_hetero* is still a time-homogeneous and stationary process but the stationary frequency distribution is no longer shared across sites and is correlated to the site-specific rate of evolution.

In both evolutionary models, twenty rate categories are used for the discrete gamma model. The gamma shape parameter α is set to 0.5, which is a reasonable value for the RNA genes being studied. This corresponds to an L-shaped distribution with most of the sites evolving slowly (see section 5.2). The homogeneous frequency parameters of *S_homo* are $\{\pi_A, \pi_C, \pi_G, \pi_T\} = \{0.40, 0.25, 0.15, 0.20\}$. Base frequency parameters used for each gamma category in *S_hetero* are given in table 5.1 and are shown in Figure 5.3. In *S_hetero*, fast-evolving sites are G+C rich and A+T poor compared to slow evolving sites. Note that the frequency parameters for each gamma category are chosen so that the frequency distribution, averaged over the whole sequence, is that given for *S_homo*. Since *S_hetero* is still a time-homogeneous and stationary process, this average composition is expected to remain constant over time and it is also expected that contemporary and ancestral sequences exhibit the pattern of compositional variation across sites shown in Figure 5.3. Exchangeabilities are set at the same reasonable values in both evolutionary models: $\rho_{CT} = 2.5$ and $\rho_{transversion} = 0.4$. Since the "rate ratios" parameterization is used in this chapter, the reference ρ_{AG} is equal to 1.0.

Sequences are then evolved along the branches of an arbitrary 10-species tree which is common for both models (see Figure 5.2).

rate category	substitution rate	π_A	π_C	π_G	π_T	π_{C+G}
1	0.0013	0.4509	0.1875	0.0839	0.2778	0.2714
2	0.0092	0.4468	0.1957	0.0901	0.2674	0.2858
3	0.0251	0.4399	0.2035	0.0981	0.2585	0.3016
4	0.0493	0.4327	0.2096	0.1077	0.2500	0.3173
5	0.0821	0.4256	0.2150	0.1171	0.2423	0.3321
6	0.1242	0.4187	0.2198	0.1258	0.2357	0.3456
7	0.1763	0.4135	0.2241	0.1330	0.2294	0.3571
8	0.2394	0.4086	0.2297	0.1395	0.2222	0.3692
9	0.3150	0.4038	0.2349	0.1458	0.2155	0.3807
10	0.4047	0.4003	0.2405	0.1514	0.2077	0.3919
11	0.5111	0.3961	0.2465	0.1573	0.2000	0.4038
12	0.6375	0.3934	0.2524	0.1622	0.1920	0.4146
13	0.7883	0.3893	0.2592	0.1681	0.1835	0.4273
14	0.9704	0.3858	0.2666	0.1733	0.1742	0.4399
15	1.1940	0.3819	0.2751	0.1786	0.1645	0.4537
16	1.4757	0.3772	0.2840	0.1841	0.1547	0.4681
17	1.8455	0.3716	0.2947	0.1894	0.1443	0.4841
18	2.3652	0.3646	0.3057	0.1941	0.1355	0.4998
19	3.2037	0.3564	0.3191	0.1978	0.1266	0.5169
20	5.5820	0.3429	0.3364	0.2027	0.1180	0.5391

Table 5.1: Frequency parameters used to generate replicates with *S_hetero*.

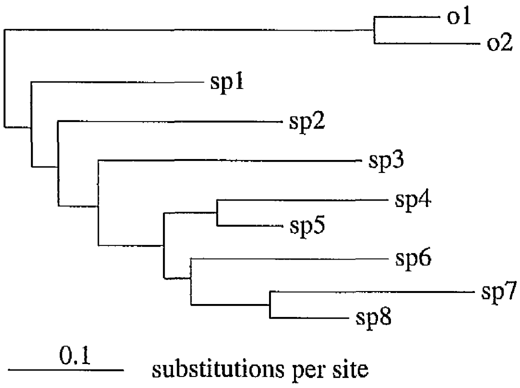


Figure 5.2: The tree that generated the synthetic datasets used in this chapter.

Replicates of *S_hetero* and *S_homo* were used to test the empirical Bayesian method described previously. Ten synthetic alignments (20, 000 nucleotides each) were generated with *S_hetero* and subsequently analyzed. During the analysis, the true topology, which is known in this case, was assumed. A "pattern-heterogeneous" TN93 substitution process modelling rate heterogeneity with twenty discrete gamma categories was used. This evolutionary model was consequently identical to *S_homo* but true parameter values were replaced by MLEs during the analysis. This model was able to model ASRV, but could not account for the fact that the composition was not constant across sites.

Figure 5.3 shows results of the empirical Bayesian method when it was applied to the first replicate. Results emphasize the fact that the method can still be used to detect compositional variation across sites even if the evolutionary model used does not model it. Nevertheless, results also indicates that the method might underestimate the extent of variation when it exists. Results with other replicates were similar and are not shown.

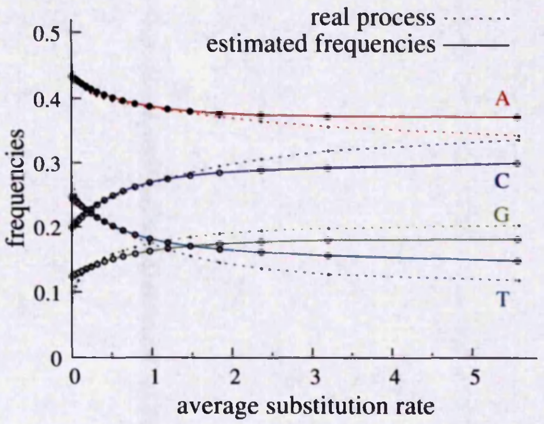


Figure 5.3: Compositional variation across sites: The frequency parameters found by the empirical Bayesian methods for each gamma category are contrasted with the real frequency parameters, which are known in this case.

Ten replicates of *S_homo* (only 1,000 nucleotides each) were also generated and analyzed with the empirical Bayesian method. The same homogeneous TN93 substitution model with 20 discrete gamma rate categories was used during the analysis. For the ten replicates, frequency parameters estimated for each rate category were found to be quite close to the stationary distribution, which is also homogeneous across sites. Differences between the estimated composition for each category and the real uniform frequencies were in the tight range [-2.48%,+2.56%] (95% confidence interval). Moreover, no significant trends related to the substitution rate were visible. This suggests that the method is not finding heterogeneity when there is none, even with relatively small datasets.

5.3.3 Application with real RNA genes of 13 primates

The ability of the empirical Bayesian method to estimate compositional variation patterns has then been demonstrated and the method was consequently applied to an empirical case. Compositional trends, and their correlation with the site-specific rate of evolution, were studied in the mitochondrial RNA genes of 13 primates (gorilla, human, chimpanzee, pygmy chimpanzee, orangutan, gibbon, baboon, barbary ape, capuchin, loris, tarsier, ring-tailed lemur, malayan flying lemur). Three species from the grouping Laurasiatheria were used as an outgroup (dog, cow and rhinoceros). The genes used are the complete set of mitochondrial tRNAs and rRNAs. This dataset was extracted directly from the dataset used for mammalian phylogenetic inference by Hudelot et al. (2003) and sequence accession numbers can be found in this article.

The consensus secondary structure is a key part of this alignment since nucleotides are treated differently according to their position. A TN93 substitution model was used with RNA loops and the 7D model was used with RNA stems. The gamma distribution of rates was approximated with eight discrete gamma categories in both blocks. The empirical method requires us to assume a specific topology and the majority-rule consensus topology found by a Bayesian analysis of this dataset was used. The chosen tree contains dubious clades, *e.g.*, the unlikely gorilla-human sister relationship found in Hudelot et al. (2003), but this is of little consequence (other topologies were tried and results were consistent).

Graphs for the estimated composition of RNA loops and RNA helices are shown in Figure 5.4. The estimated frequency distribution for each discrete rate category is plotted against the average substitution rate of the category. In RNA stems, frequencies of symmetrical pairs (*e.g.*, $\pi_{G:U}$ and $\pi_{U:G}$) were summed for clarity. In loops, π_G appears to be negatively correlated with the rate of evolution and this nucleotide is underrepresented at fast evolving sites. In RNA stems, a striking increase in correlated with the site-specific evolutionary rate is observed for $\pi_{A:U+U:A}$ and π_{MM}, whereas $\pi_{G:C+C:G}$ is decreasing.

The high **A** content observed in unpaired regions of primate mitochondrial RNA genes is explained by Gutell et al. (2000) who suggest that unpaired **A**

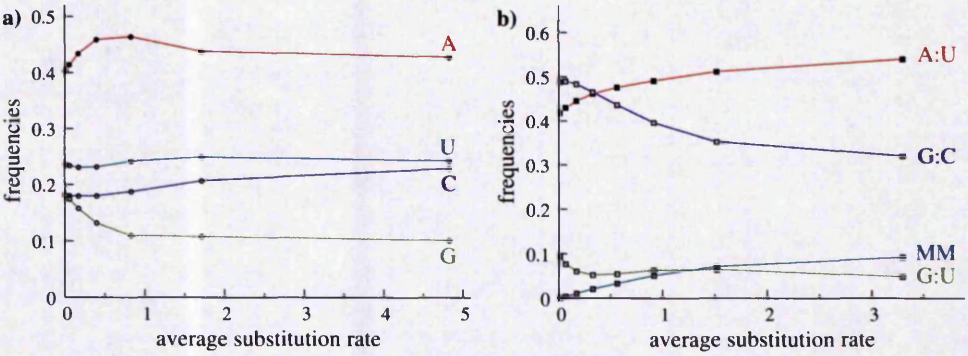


Figure 5.4: Compositional variation across sites with mitochondrial RNA genes, (a)loops and (b)helices. Note that the 7D model distinguishes the pairs **X:Y** and **Y:X** but state frequency parameters were summed to produce the curves in (b).

nucleotides are crucial components for the formation of three-dimensional rRNA molecules. The biological processes responsible for other compositional trends are hardly understood but it appears that the observed variations can mostly be explained by the combined effects of the mutation and selection processes. Fast categories are under reduced selection and respond to the mutational pressure whereas the frequency distribution observed at sites under purifying selection reflects an average of the site-specific selection pressure over all slow evolving sites. This fits nicely with the results. In loops, the decreasing trend in **G** content is certainly due to the strong mutational bias away from **G** in mammalian mitochondrial genes. This mutational bias is often invoked to explain the low **G** content at the third codon position for mitochondrial protein-coding genes found on the H-strand (Gibson et al., 2005) and the results here are consistent with the hypothesis of the deamination of **C** on the H-strand which results in the decrease of **G** and increase of **A** in RNA products (Reyes et al., 1998). Indeed, the two rRNA genes, which account for two-thirds of the dataset, are on the H-strand. Moreover, when the original dataset was split in two, to separate tRNAs and rRNAs and to study independently compositional variation across sites in these genes, the decreasing trend in **G** content was much more striking for rRNAs than for tRNAs. This result was expected since tRNA genes are found both on the L and the H strands (curves not shown). Nevertheless, since a consensus secondary structure was used, one cannot exclude the possibility of a "contamination" of the loop partition with slow evolving paired sites, more G+C rich than the standard composition in RNA loops. However, one would also expect strong compositional variations for the **C** and **U** bases in such a case. Since **G** and **A** are the only unpaired nucleotides exhibiting trends, this explanation seems less plausible.

Compositional trends in primate mitochondrial RNA helices are stronger and also fit with what could reasonably be expected. The frequency of mismatches increases steadily with the evolutionary rate, reflecting a weaker selection pressure to maintain the RNA secondary structure at fast evolving sites. Since **G:C** pairs are thermodynamically stronger than **A:U** pairs and have a strong stabilizing effect on the structure of RNA molecules, it seems reasonable that slow evolving regions, subject to a stronger selection pressure, are **G:C** rich compared to fast-evolving regions. Nevertheless, the mutation bias away from **G** mentioned previously might also be responsible for the striking decrease in the **G:C** content in favor of the **A:U** content. More studies are probably needed to assess how ubiquitous the decrease in the **G:C** content is.

5.3.4 Discussion

This study is limited to the correlations between site-specific evolutionary rate and composition in RNA genes. Nevertheless, similar trends were also found with protein-coding genes at the nucleotide/codon level and at the amino-acid level[1]. With these genes, mutational biases are acting at the nucleotide level whereas selective constraints are acting at the codon level. This gives rise to unexpected patterns of compositional variation across sites.

The empirical Bayesian method proposed for the study of compositional variations among RNA sites suffers from some drawbacks. Mayrose et al. (2004) reported that Bayesian methods perform quite well for the estimation of the site-specific substitution rate but the accuracy of this approach is still limited. As already emphasized, one important issue is that the method assumes compositional homogeneity across sites to classify sites according to their evolutionary rate and that is probably not appropriate when the objective is to study spatial frequency variations. Results with *S_hetero* (Figure 5.3) suggest that the use of a pattern-homogeneous model and/or the use of rate category posterior probabilities as weights tend to flatten the resulting curves. Consequently, no claim is made concerning the accuracy of the curves plotted with this method since they

[1] A significant part of this work was done by Antoine Buxerolles during his MSc project.

are most likely underestimating the extent of spatial compositional variation in real data. However, in spite of the fact that the method might not be reliable enough to quantify compositional variation numerically, put together with the negative results obtained with dataset generated by *S_homo*, it is concluded here that the striking trends exhibited with real sequences are genuine and that spatial compositional variation is a common phenomenon that can be correlated with the site-specific rate of evolution.

5.4 Effects on standard phylogenetic inference methods

Since most genes used in phylogenetic inference are subject to selective constraints that vary quantitatively along the length of the sequences, variation of nucleotide composition across sites is more likely to be the norm than the exception. The biasing effects of compositional heterogeneity across sites on phylogenetic methods that do not account for it is investigated here

5.4.1 Simulations

To assess the impact of compositional heterogeneity across sites on phylogenetic estimates, one hundred replicates were generated using the evolutionary model *S_hetero* ($20,000$ sites per alignment) and subsequently analyzed with standard homogeneous methods. For each replicate, branch lengths and substitution model parameters (*i.e.*, frequencies, exchangeabilities and gamma shape parameter) were reestimated by ML optimization assuming the tree topology which is known in this case. The inference model was a TN93 substitution model and ASRV was accounted for with twenty discrete gamma rate categories. The inference model is consequently identical to the generative model *S_homo* but it cannot account for the variation of composition across sites present in replicates generated by *S_hetero*.

The following values were recovered for the frequency parameters: $\{\pi_A = 38.7 \pm 0.5\%, \pi_C = 27.3 \pm 0.5\%, \pi_G = 16.3 \pm 0.4\%, \pi_T = 17.7 \pm 0.4\%\}$. These MLEs are clearly biased towards the composition of fast evolving sites (given in table 5.1). Even though the site-specific composition depends on the rate category in these synthetic datasets, one would have expected the MLEs for the frequency parameters to be close to the stationary nucleotide distribution when it is averaged over the whole sequence: $\{\pi_A = 40\%, \pi_C = 25\%, \pi_G = 15\%, \pi_T = 20\%\}$. The bias also exists when the inference model does not account for ASRV but it was less noticeable: $\{\pi_A = 39.2 \pm 0.5\%, \pi_C = 25.4 \pm 0.4\%, \pi_G = 15.8 \pm 0.4\%, \pi_T = 19.6 \pm 0.4\%\}$.

Frequencies were not the only parameters affected. Yang et al. (1994) and Huelsenbeck and Nielsen (1999) noticed that branch lengths were underestimated when using a simpler evolutionary model that does not accommodate rate heterogeneity or transition/transversion rate variation across sites. Similarly, it was noticed here that all branch lengths were slightly underestimated (by 3% approximately) when variation of frequencies across sites was not accounted for. The estimation of exchangeability parameters was also affected, with $\rho_{CT} = 2.02 \pm 0.15$ and $\rho_{transversion} = 0.37 \pm 0.02$ instead of $\rho_{CT} = 2.50$ and $\rho_{transversion} = 0.40$.

These experiments were repeated with replicates generated by *S_homo* (same alignment size, 20,000 sites) to confirm that deviations from the expected result were only due to the compositional variation across sites. When replicates of *S_homo* were used, MLEs were close to their true values which was expected because ML is consistent and asymptotically efficient when the generative model and the inference model are the same. The alignment size was large enough to grant these results. Since the different biases highlighted above are not observed with replicates generated by *S_homo*, compositional variation across sites in *S_hetero* must be responsible for them.

Above, results are presented in which the frequency distribution is a free parameter of the model that is inferred with the other evolutionary parameters during ML optimization. As already mentioned, frequency parameters can directly be approximated by the empirical composition of contemporary sequences since stationarity is assumed. Previous experiments were consequently repeated when the frequency parameters are fixed to their observed empirical values. Obviously, frequency parameters were not biased anymore in this case but it was found that the other biases were more pronounced in such a setting. Branch lengths were even more underestimated (by 4% approximately) as were the exchangeability parameters ($\rho_{CT} = 1.85 \pm 0.15$ and $\rho_{transversion} = 0.36 \pm 0.02$).

Effects on topology estimates were studied in the Bayesian framework. One hundred alignments were generated using *S_hetero* with 1,000 sites per replicate and analyzed with the MCMC sampler of **PHASE**. Once again, the substitution model used during the inference was similar to the generative substitution model of *S_homo* (TN93, twenty gamma rate categories), but branch lengths and substitution parameters were treated as unknown variables. The tree topology was also considered as a free parameter here and standard priors were used. The BPPs that measure clade supports were compared to the BPPs found when compositional variation across sites is accounted for by the substitution model. To perform such a comparison, experiments were repeated using an inference model which is similar to *S_hetero*. The tree topology and the substitution parameters were still considered as unknown parameters but the base frequencies used for the 20 rate categories were not assumed to be homogeneous and were fixed to their true values. On top of the biases highlighted above in the ML framework, it was found that the BPPs used to measure the support for different clades were sensitive to the model used. Focusing on the clades with BPPs between 50% and 95% when the true model was used (the values that are usually reported on a consensus tree), corresponding BPPs obtained with the spatially homogeneous model were slightly, but significantly, different (±5% on average, and ±10% to 20% in some cases).

5.4.2 Discussion

The bias towards the composition observed at fast evolving sites can easily be explained. When a single stationary distribution of frequencies is assumed and shared across sites evolving at different rates, an invariant site provides only one single independent sample of this distribution whereas, at the other extreme, a site with infinite substitution rate would provide N uncorrelated samples, N being the number of taxa in the alignment. Deviations of frequency estimates from the empirical composition observed at fast evolving sites consequently have a larger detrimental effect on the likelihood and, as confirmed by the results with simulated datasets, ML and Bayesian inference methods favor frequency parameters that are closer to the composition of the fast evolving sites. This was not tested here but one can reasonably expect the divergence between empirical and estimated frequencies to grow as taxon sampling increases because N will increase.

Although the effects of compositional heterogeneity across sites on the estimation of substitution parameters are worrying, the impact on branch lengths and topology estimates seems quite limited. Nevertheless, results suggest that it is not completely negligible. Disregarding pattern-heterogeneity of the substitution process has probably little effect on the inferred phylogeny in general but differences can appear when using a substitution model that accounts for it.

5.5 Modeling compositional heterogeneity across sites

The evolutionary processes that cause the observed compositional trends with respect to the site-specific evolutionary rate are poorly understood and, in any case, hard to model explicitly. One simple method would be to use different equilibrium frequencies for each rate category and to consider each frequency vector as a free parameter to be estimated from the data. Nevertheless, this method would be parameter-rich which is an issue both statistically and computationally. Moreover the number of gamma categories used would have a direct impact on the substitution model and the number of parameters, which would also be annoying because discretization of the gamma distribution is primarily intended to be a mathematical convenience. In spite of these problems, such a substitution model was attempted. When tested with synthetic datasets, generated by evolutionary models similar to *S_hetero*, it was found that the method was not able to fit the real frequency curves of the generative model. Equilibrium frequencies at slow evolving sites were poorly recovered and highly variable.

5.5.1 A Gaussian Process model

Empirical results suggest that the composition does not vary strongly between neighboring rate categories. Consequently, it seems reasonable to solve the issues with the previous unconstrained method by assuming that equilibrium frequencies vary smoothly with respect to the site-specific evolutionary rate. Simple parametric models that express frequencies as a function of the substitution rate were tested but, once again, the fit to the real frequency curves was often found to be unsatisfactory. These parametric models failed badly with some datasets (work not shown).

Parametric methods failed because they are too rigid. Instead, a Gaussian process can be used to incorporate the prior of smoothness into the model of compositional variation. Gaussian processes are becoming popular in the machine learning community (Rasmussen and Williams, 2006) and they have already been applied in various fields, usually for classification (Gibbs and MacKay, 2000) and regression problems (Williams and Rasmussen, 1996). See also Chu et al. (2005) for a recent application in computational biology. Easy to implement and to interpret, Gaussian processes are also appealing because their nonparametric nature does not constrain the model to a specific functional form. In the current application, Gaussian processes can simply be considered as useful smoothing devices that return a prior probability of observing a collection of frequency vectors without making excessively strong assumptions on the underlying causes of the variation across sites. Frequencies at each rate category are treated as free parameters but strong variations are penalized by the Gaussian process prior. The method was implemented in the Bayesian framework using MCMC techniques.

Gaussian processes are fully defined by their mean and covariance function. The covariance matrix $\mathbf{C}$ specifies how similar the composition is between each pair of rate categories. In phylogenetic inference, substitution rates are always expressed relative to each other and consequently the problem is reparameterized with $x_c = \log(r_c)$, where r_c is the average substitution rate of the rate category c. The following covariance form is used:

$$C(x_i, x_j) = \theta_0^2 \exp\left\{ -\theta_1^2 (x_i - x_j)^2 \right\} + \theta_2^2 + \theta_3^2 x_i x_j + \delta_{ij}\theta_4^2 \quad . \tag{5.4}$$

The hyperparameter θ_0 defines the amount of variation expected for a typical function, θ_1 scales the rate abscissa, θ_2 and θ_3 introduce a linear trend in the Gaussian process and define the mean expected values of the process as a straight line with respect to the evolutionary rates x while θ_4 is a jitter element on the diagonal of the covariance matrix which prevents it from being ill conditioned. A distinct covariance matrix $\mathbf{C}_\mathrm{i}$ (*i.e.*, a specific vector $\Theta_i = \{\theta_0, \theta_1, \theta_2, \theta_3, \theta_4\}$) is used for each of the n possible states of the model (*e.g.*, four nucleotides for TN93 and seven pairs for 7D) defining a set of hyperparameters $\mathbf{\Theta} = \{\Theta_1, \Theta_2, \ldots, \Theta_n\}$. A unique (arbitrary) prior is used on each Θ_i: $\theta_0 \sim Ex(10)$, $\theta_1 \sim Ex(1)$, $\theta_2 \sim$

$Un(0, 10)$, $\theta_3 \sim Un(0, 10)$, $\theta_4 = .005$.

For each model, the n frequency parameters used at a specific rate category are reparameterized with n "activation" values using a softmax function to remove the usual constraints imposed on frequency parameters: $0 \leq \pi_i \leq 1$ and $\sum_i \pi_i = 1$,

$$\pi_i(r) = \frac{\exp a_i(r)}{\sum_j \exp a_j(r)} \quad . \tag{5.5}$$

The $a_i(x)$ vectors are not uniquely defined. It would have been possible to keep the extra degree of freedom for each gamma category and to let the Gaussian process prior resolve the identifiability issue. Instead, the simple constraint $\forall r, \sum_j a_j(r) = 0$ was added although it has some effects on the prior for the variations. This constraint was imposed to limit the correlations between the n activation values at each rate category and improve the mixing behaviour.

The probability density of a set of activations $\mathbf{a_i} = (a_i(r_1), \ldots, a_i(r_k))$, where $\{r_1, \ldots, r_C\}$ are the rates of the discrete gamma categories is:

$$p(\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n} | \Theta_1, \Theta_2, \ldots, \Theta_n) = \frac{1}{Z} \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} \mathbf{a_i}^T \mathbf{C_i}^{-1} \mathbf{a_i}\right\} \prod_{c=1}^{C} \delta(\sum_{i=1}^{n} a_i(r_c)) \quad , \tag{5.6}$$

where $\mathbf{C_i}(x, y) = C_i(\log(r_x), \log(r_y))$, Z is a normalizing factor (see appendix B) and $\delta(x)$ is the Dirac delta function.

When a standard substitution model is used for Bayesian inference in **PHASE**, a Dirichlet proposal mechanism is used to propose new equilibrium frequencies from the current distribution vector during a MCMC run (see section 3.4.1). Multiple frequency vectors are used in the substitution model presented here and they are updated independently using the same mechanism. Since strong correlations between the frequencies of neighboring categories have been introduced, distant moves are now regularly refused and the parameters that control the spread of the Dirichlet proposals are initially chosen (and also tuned during the MCMC "burn in" period) to ensure a reasonable acceptance rate when sampling from the chain. Since large moves are not possible, mixing properties are not impressive and a huge number of cycles are necessary to obtain a reliable sample from the stationary distribution of each gamma category. A possible solution would be to design a suitable move affecting all of the frequencies at once in order to preserve the correlations. For instance, it should be possible to design a proposal that would change a set of frequency parameters used at a specific rate category by replacing them with random values drawn from the Gaussian process used as a prior according to the frequency parameters at other rate categories. Nevertheless, this was not attempted here since computational issues are not of greatest interest and instead very long runs were used. Two runs were performed for each simulation and we checked that the final posterior probability distributions of frequency parameters were similar for each gamma category.

5.5.2 Test with simulated datasets

The new model that constrains compositional variations between neighboring rate categories with a Gaussian process was tested with datasets generated by *S_homo* (1,000 nucleotides) and *S_hetero* (20,000 nucleotides), the two evolutionary models introduced in section 5.3. Rate variation across sites (with *S_homo*) and rate+compositional variations across sites (with *S_hetero*) were simulated using twenty discrete categories in both cases but, for computational reasons, the TN93 substitution model used during the inference accounts for rate+compositional variations across sites with only eight discrete categories. In Figure 5.5, the twenty sets of frequency parameters of the generating evolutionary model *S_hetero*, given in table 5.1 and previously shown in Figure 5.3, are compared with the eight sets of frequency parameters inferred when the Gaussian process model is used. Mean posterior estimates (MPEs) were used to plot the curves and are accompanied with corresponding 95% credibility intervals. The results show excellent agreement between inferred and true frequencies. MPEs for the hyperparameters Θ (see equation (5.4)) are given in table 5.2. In the same figure, a 1,000 nucleotides long alignment, generated with *S_homo*, is used to perform a negative control. Although some trends were recovered, the real uniform frequency parameters lies within the credibility intervals.

5.5.3 Application to the primate dataset

The new model that constrains compositional variations between neighboring rate categories with a Gaussian process was also tested with real data. The primate dataset studied in the previous section was used. This dataset was analyzed using the same two substitution models as previously: TN93 for loops and 7D

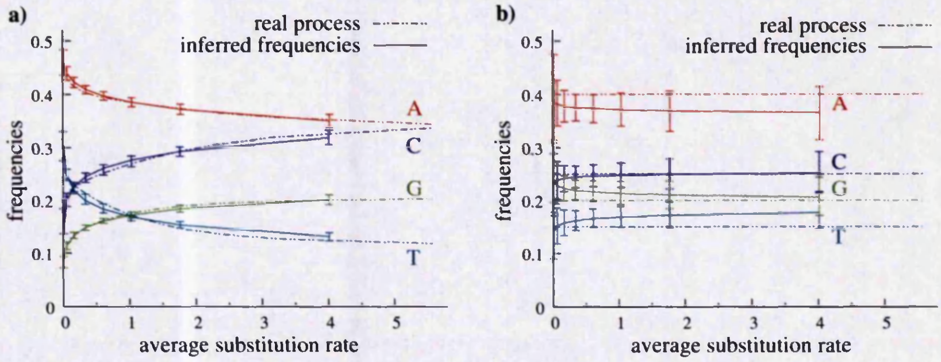


Figure 5.5: Results with the GP model that accounts for compositional variation across sites. Mean posterior estimates inferred for the frequency parameters of each gamma category are compared to the frequency parameters used to generate the data. (a) The alignment is 20,000 nucleotides long and was generated with *S_hetero*, equilibrium frequencies vary across sites and depend on the site-specific evolutionary rate. (b) The alignment is 1,000 nucleotides long and was generated with *S_homo*, equilibrium frequencies of the generating process are uniform. The 95% credibility intervals around the mean posterior estimates are shown. Twenty discrete gamma categories were used to generate the sequences but only eight were used during the inference and therefore the discrete gamma model for the real process spans a larger range of rates.

	θ_0	θ_1	θ_2	θ_3
TN93 - **A**	0.0569	0.4121	3.0336	1.8897
TN93 - **C**	0.0693	0.4270	2.3330	2.1618
TN93 - **G**	0.0570	0.4245	2.7306	2.1654
TN93 - **T**	0.0741	0.3885	2.7612	2.2992

Table 5.2: Mean posterior estimate for the hyperparameters of the Gaussian process prior: inferred values for θ_0, θ_1, θ_2 and θ_3 for each state (see equation (5.4)). Results are given for a dataset generated with *S_hetero* (20,000 nucleotides).

for stems. The eight gamma categories assumed in both models were assigned different sets of frequencies and Gaussian process priors were used in both models to avoid strong variations at neighboring rate categories.

Figure 5.6 shows the MPEs for the frequency parameters at each rate category accompanied with 95% credibility intervals. Inferred frequency parameters are compared with the frequency parameters found with the Bayesian empirical method in the previous section. The fit is less impressive than with simulated data (although reasonable) but one must keep in mind that the empirical estimates probably flatten the curves and underestimate the variation of frequencies across sites (as previously observed in figure 5.3).

MPEs for the hyperparameters $\boldsymbol{\Theta}$ (see equation (5.4)) are given in table 5.3 for both substitution models.

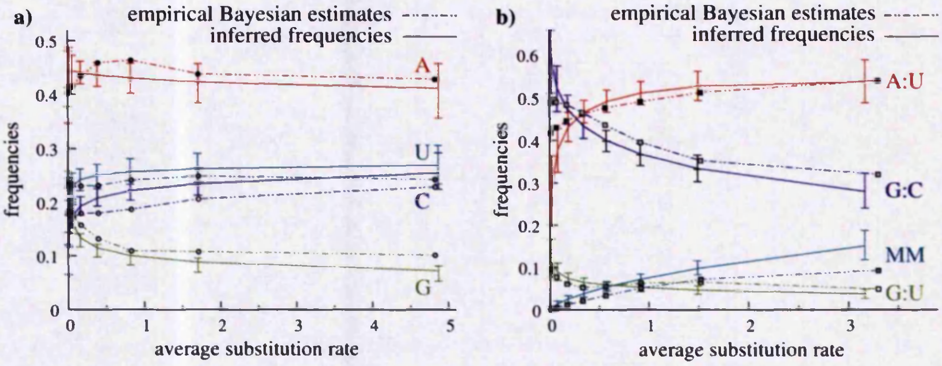


Figure 5.6: Results with the GP model that accounts for compositional variation across sites. The primate dataset used is, once again, partitioned into two blocks: (a) loops and (b) stems. Mean posterior estimates inferred for the frequency parameters of each gamma category are compared with the results of the empirical Bayesian method described in the previous section.

	θ_0	θ_1	θ_2	θ_3
TN93 - **A**	0.0439	0.4438	3.3272	1.2966
TN93 - **C**	0.0419	0.4512	1.8451	2.2112
TN93 - **G**	0.0491	0.4180	3.5620	2.3813
TN93 - **T**	0.0476	0.3929	2.3093	2.0194
7D - **A:U**	0.0575	0.4599	3.7055	2.3437
7D - **G:U**	0.0607	0.5238	4.0645	2.3712
7D - **G:C**	0.0649	0.4538	3.2261	2.5534
7D - **U:A**	0.0643	0.4626	3.6084	1.9333
7D - **U:G**	0.0596	0.4912	4.1492	2.6997
7D - **C:G**	0.0631	0.4561	3.1251	2.1755
7D - **MM**	0.0606	0.4663	2.4701	3.2818

Table 5.3: Mean posterior estimate for the hyperparameters of the Gaussian process prior: inferred values for θ_0, θ_1, θ_2 and θ_3 for each state (see equation (5.4)).

5.5.4 Discussion

In Thorne et al. (1996) and in subsequent works from the Goldman group (*e.g.*, Goldman et al. (1998); Liò and Goldman (2002)), a trained HMM model is used to incorporate information on the variation of substitution patterns across sites before the inference. On the contrary, in more recent works (Pagel and Meade, 2004; Lartillot and Philippe, 2004; Soyer et al., 2002), evolutionary models that could potentially detect and characterize new patterns of heterogeneity across sites with little *a priori* knowledge were used. Even though different substitution models are used for each rate category, we have to concede that we did not completely relax the "one model fits all" assumption of standard homogeneous substitution models in this work. Incorporating strong prior constraints to prevent variations at neighboring rate categories clearly prevents the putative patterns of evolution to emerge freely but it is recalled that a completely unconstrained mixture model is not a practical option. As pointed out at the beginning of section 5.5, such a model would over-fit.

The mutational process, which is not to be confounded with the substitution process, can be assumed constant across sites and it seems reasonable to suppose that frequencies at fast evolving categories are correlated. Nevertheless, one can also reasonably argue that slow evolving sites should not be grouped according to their average substitution rate but rather according to the site-specific selection pressure which drives the process equilibrium frequencies at these sites. While that point has some merit, it should be remembered that the new model proposed here is primarily designed for RNA helices. Since structural stability is the common and predominant selection factor, it is hypothesized that slow evolving pairs are under similar selection pressure and can be treated with the same process .

Results using a Gaussian process prior on the primate dataset and the synthetic datasets are promising. With the primate dataset, inferred frequencies do not seem to fit very well the empirical curves (figure 5.6) but, as previously mentioned, results from the empirical Bayesian methods are just a best guess derived from a homogeneous model and the actual amount of variation might be seriously underestimated. Results with *S_hetero* (figure 5.5) are quite impressive because the alignment is of sufficient size and because the frequencies of the *S_hetero* process were close to linear in the logged substitution rate scale, which is an important assumption of the prior. Sensitivity and effect of the prior should always be investigated in Bayesian inference. In this case, frequencies inferred at slow evolving sites are clearly more influenced by the trend imposed by the prior rather than the data. A low value for the hyperparameter θ_0 is important to prevent unrealistically large variations at slow evolving sites similar to the variations that would be observed by using the completely unconstrained mixture model that does not attempt any smoothing. This approach to choosing the prior might seem rather ad-hoc. Although the functionality was not used here, a hierarchical scheme could have been used and the user is allowed to define complex parameterized prior distributions for θ_0, θ_1, θ_2 and θ_3. With some datasets we found that the method would not behave properly unless θ_0 is constrained close to zero with an exponential prior. One obvious reason might be that patterns uncorrelated to the evolutionary rate were present in these real datasets and negating the attempt to smooth variations. However, time-heterogeneity and "difficult" species seem a more likely explanation for these occasional failures.

5.6 Effects on time-heterogeneous methods

In section 5.4, it was shown that compositional heterogeneity across sites has significant effects on estimates from standard phylogenetic methods. The time-heterogeneous methods introduced in chapter 4 allow for compositional variation over time but, except for ASRV, they also assume that evolutionary patterns are constant across sites. Similar problems are consequently expected to arise. Since the unique stationary equilibrium distribution of standard time-homogeneous methods was found to be biased towards the composition observed at fast evolving sites, it can be predicted that the collection of equilibrium frequencies used by time-heterogeneous models on the branches of the phylogeny are also biased towards the composition at fast evolving sites. This is demonstrated here. It is also shown that the estimated ancestral state distribution is, on the contrary, biased towards the composition observed at slow-evolving sites. This casts some doubts on the reliability of the ancestral frequency estimates returned by time-heterogeneous methods and could explain why the results in section 4.7 are different from the results of the original study by Galtier et al. (1999).

5.6.1 Simulations

To analyze the effect of compositional variation across sites on time-heterogeneous methods, the program **EVAL_NH**, from the **NHML** package (Galtier and Gouy, 1998), was used. This program is designed to optimize the branch lengths and substitution parameters of a time-heterogeneous model when the topology is known. It was chosen because it implements the GG98 substitution model (Galtier and Gouy, 1998) which has few parameters and is consequently easily tractable.

Replicates of *S_hetero* (100 alignments of 20,000 nucleotides) were generated and analyzed with **EVAL_NH**. The true topology was assumed once again. Experiments were performed with and without a discrete gamma model. Eight gamma categories were used in the first case. When the time-heterogeneous and rate-heterogeneous evolutionary model was used, the inferred ancestral G+C frequency ω was 33.85% (95% CI:[33%, 34.57%]). The bias was less striking but still visible when the inference model did not account for rate-heterogeneity (ω=37.3%, 95% CI:[36.56%, 38.06%]). The evolutionary model *S_hetero* is admittedly heterogeneous across sites but it is still time-homogeneous and stationary. Consequently, the ancestral G+C frequency in replicates of *S_hetero* was expected to be approximately equal to the average stationary frequency ($\pi_G + \pi_C = 40\%$). The recovered estimate of ω is clearly biased towards the G+C frequency of slow evolving sites ($\pi_G + \pi_C = 27.1\%$ for the slowest gamma category).

The GG98 model is a time-heterogeneous version of the T92 model (Tamura, 1992). Since the T92 substitution model is less complex than the TN93 model used to generate synthetic datasets in *S_hetero*, one might object that the violation of these assumptions is responsible for the observed differences. Nevertheless, similar results were reproduced when the generative TN93 substitution model was replaced with a T92 model (work not shown). Moreover, when the experiments were repeated with replicates generated by *S_homo*, which is homogeneous across sites, results confirmed that using a T92 model instead of a TN93 model had only a negligible impact on the results and the ancestral G+C frequency recovered was much closer to its real value: $\omega = 38.8 \pm .7\%$ with rate heterogeneity among sites and $\omega = 39.7 \pm .8\%$ without. As predicted, it was also found that the equilibrium G+C frequency distributions estimated on different branches of the tree were biased towards the G+C frequency at fast evolving sites and higher than the expected 40% (result not shown).

5.6.2 A mesophilic LUCA?

It has just been demonstrated that compositional variation across sites could have a very detrimental effect on the ancestral frequency estimates returned by time-heterogeneous methods. When analyzing replicates of *S_hetero* with **EVAL_NH**, the ancestral G+C content was underestimated by approximately 6%. Although such a difference might seem small, this potentially has grim consequences for the results presented in section 4.7 and for the study of Galtier et al. (1999). Indeed, the difference between the G+C content of mesophilic and thermophilic species is of such an order of magnitude (see Figure 4.12).

Di Giulio (2000) reported that Galtier et al. (1999)'s results could not be repeated using maximum parsimony and the bias reported here is a likely explanation for this disagreement. Compositional heterogeneity across sites could have had a significant impact on the ancestral G+C content proposed by Galtier et al. (1999) and this hypothesis is tested here. Since the possible bias of nonhomogeneous methods is related to a directional trend in the composition with respect to the site-specific evolutionary rate, variations of the G+C content across sites were studied in the Tree of Life dataset introduced in chapter 4.

To this end, the Bayesian empirical method introduced in section 5.3 was used. It was assumed that the tree topology shown in figures 4.10 and 4.11 was correct. It is recalled that the method used to estimate the correlation between composition and site-specific evolutionary rate is based on a time-homogeneous model and results are consequently not dependent on the root position. To evaluate the potential impact of compositional heterogeneity across sites on Galtier et al.'s ancestral G+C estimate, the complete Tree of Life dataset (*i.e.*, loops+stems) was analyzed with a T92+dG20 model. To evaluate the potential impact on the ancestral G:C+C:G estimate proposed in chapter 4, stems were analyzed with a 7D+dG20 model. Columns with gaps and ambiguous nucleotides/pairs were removed before the analysis in both cases.

Results are presented in Figure 5.7. At the top of the figure, the two ancestral compositions, one for each possible rooting, are compared to the empiral composition found in present-day rRNA sequences. The analysis on the top-left was performed with **EVAL_NH** and the ancestral G+C contents are consequently ML estimates. Naturally, the results presented here are highly consistent with the results of in Galtier et al. (1999) since the two analyses were similar. On the top-right, results found in chapter 4 are reproduced, the ancestral G:C+G:C contents in stems are mean posterior estimates. At the bottom of the figure, results obtained with the empirical Bayesian method are given. The G+C content in loops and stems (bottom-left) is positively correlated with the site-specific rate of evolution. This suggests that the ancestral G+C content initially proposed by Galtier et al. (1999) was an underestimate. The G:C+C:G content in stems (bottom-right) seems weakly correlated to the site-specific rate of evolution but the trend is not really clear. It is possible that the ancestral G:C+C:G contents proposed in this thesis are overestimated, but the relationship is nonmonotonic so it is not clear.

5.6.3 Discussion

When analyzing replicates of *S_hetero* with **EVAL_NH**, the ancestral G+C estimate was found to be strongly biased towards the frequencies observed at slow evolving sites. The Bayesian time-heterogenous method developed in chapter 4 certainly suffers from the same problems. The bias has probably a limited effect on the accuracy of the phylogenetic reconstruction and the results presented in this last section should not be taken as a plea for avoiding the use of time-heterogeneous methods. Compositional heterogeneity in time is certainly as ubiquitous as compositional heterogeneity across sites and using a time-heterogeneous method is more likely to improve phylogenetic accuracy even when site-specific patterns of evolution are highly variable across sites. Nevertheless, impacts on the inferred ancestral composition are far from negligible and studies that focus on the estimation of the ancestral DNA composition should be complemented by a study of the correlation between nucleotide frequencies and site-specific evolutionary rates to confirm whether results can be trusted.

It has been shown that site variations in selection pressure could bias a time heterogenous method. One could equally argue that time-heterogeneous evolution would bias a site-heterogenous model. Although this was not investigated here, one could certainly show, with similar simulations, that the equilibrium frequencies found by site-heterogeneous methods at slow evolving sites are biased

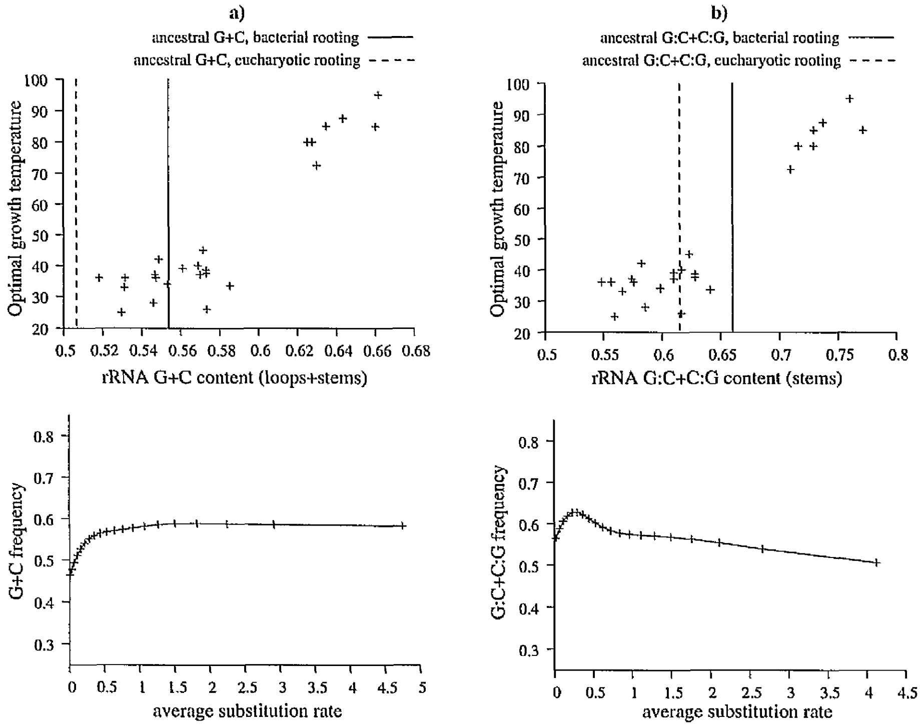


Figure 5.7: Frequency variation with respect to the average evolutionary rate in the Tree of Life dataset. LSU and SSU genes were concatenated. Top-left) the ML estimates for the ancestral G+C content returned by the time-heterogeneous GG98 model are compared with the empirical G+C content of present-day sequences. Top-right) The mean posterior estimates for the ancestral G:C+C:G content in stems returned by the time-heterogeneous 7D model implemented in **PHASE** are compared with the G:C+C:G content found in contemporary rRNA helices. Two estimates are returned in both cases depending on the rooting point. Bottom-left) G+C frequency variation with respect to the average evolutionary rate, loops and stems were joined. Bottom-right) G:C+C:G frequency variation with respect to the average evolutionary rate in stems.

towards the ancestral composition at the root of the tree and do not reflect the actual evolutionary process.

Although the sequences and the alignment used here are not exactly the same, the datasets used by Galtier et al. (1999) probably exhibited trends similar to what was shown in figure 5.7 and it is quite likely that Galtier et al. (1999) underestimated the actual ancestral G+C content of the two rRNA genes. Admittedly, results presented in this thesis do not provide strong support for a hyperthermophilic ancestor but this hypothesis could not be completely rejected. LUCA might have been a moderately thermophilic species.

Ancestral composition estimates returned by time-heterogeneous methods are certainly biased but it is unfortunately not possible to predict the amplitude of the bias. It would be tempting to approximate the amount of the underestimation or overestimation from the shapes of the curves that relate the nucleotide composition to the site-specific evolutionary rate but the number of species, shape of the tree and branch lengths certainly have a big influence on the extent of the bias. Moreover, the empirical Bayesian method used to estimate the correlation between composition and site-specific evolutionary rate is probably returning slightly innacurate results when the process of evolution is time-heterogeneous. A model that would allow simultaneously for variation in time and variation across sites seems necessary for a more accurate estimate of the ancestral composition.

5.7 Conclusion

Base composition in nuclear and mitochondrial RNA genes was found to be heterogeneous across sites. This is most likely due to variation of the selection pressure over the length of these genes and across-site heterogeneities in the substitution process. The distribution of states observed at a site was found to be correlated to the site-specific evolutionary rate and a new method was proposed to capture this aspect of sequence evolution. A Gaussian process model was used to model the variation of equilibrium frequencies with respect to the evolutionary rate and to allow for a smooth variation of the stationary distribution between neighbouring rate categories.

Most evolutionary models are pattern-homogeneous and assume that the substitution process is uniform across-sites. It is known that accounting for ASRV can tremendously improve the accuracy of phylogenetic reconstruction and it was investigated here whether the effects of compositional variation across sites could mislead standard methods. Impact on phylogenetic accuracy was found to be limited but it was shown that the equilibrium frequency parameters estimated by statistical phylogenetic methods were biased towards the composition of rapidly evolving sites. Ancestral frequency estimates returned by time-heterogeneous methods were also found to be biased towards frequency distributions of conserved sites. Caution is consequently advised when applying these methods to recover the composition of ancestral sequences. This suggests that a time-heterogeneous and pattern-heterogeneous substitution model would be most useful to recover the ancestral G+C content of LUCA's rRNA sequences accurately. Although it would theorically be possible to combine the time-heterogeneous model presented in chapter 4 with the pattern-heterogeneous model developed in this chapter, this is not (yet) a practical option because both models are computationally taxing. Combining the two would seriously worsen the situation.

Chapter 6

Conclusion

6.1 Summary of the thesis

Contemporary phylogenetic methods are no longer limited to the descriptive study of species relatedness. Current model-based approaches are as much designed to unravel the mechanisms of sequence evolution as they are to recover the evolutionary relationship of organisms. With the advent of fast MCMC techniques and the ever increasing computational power available to researchers, the trend in the past few years has been towards complex and parameter-rich models that attempt to capture a wider range of evolutionary forces. Although this is not guaranteed, it is expected that better evolutionary models increase the accuracy of reconstructed phylogeny, reduce bias, and, at the very least, do not lull the user into a false sense of overconfidence. In this thesis, complex models were built to account for the specificity of RNA genes and were implemented on the **PHASE** software package, which is freely available under the GPL license.

- The **PHASE** software, initially written by Jow (2003), was modified to allow the use of combined models. This was necessary to use complete RNA genes, *i.e.*, loops and stems, with appropriate models for each block of the partition, *i.e.*, a nucleotide substitution model for unpaired regions and a base-pair model for helices (Hudelot et al., 2003). Combined models in **PHASE** also found another use with protein-coding genes where different substitution models are used for different codon positions (Gibson et al., 2005).

- Following Foster's work (2004), time-heterogeneous substitution models were implemented using rjMCMC techniques. The focus in this thesis was on the variation of frequency parameters over time since it was known that this could mislead traditional phylogenetic methods. Nevertheless, the model implemented is much more general and could be used, for instance, to study the evolution of exchangeability parameters over time. The method is computationally tractable with reasonably sized datasets and automatically detects the amount of heterogeneity needed, in a statistically principled manner.

- It has been shown that ignoring variation of the evolutionary process across sites could have important consequences, especially with time-heterogeneous methods. Models that allow for the variation of equilibrium frequencies and exchangeability parameters across sites were therefore built. One important feature of these models, supported by empirical evidence, is that these variations are assumed to be linked to the strength of the evolutionary forces acting at a site and are consequently correlated to the variation of the site-specific evolutionary rate.

When building a complex evolutionary model, it is often tempting to add many parameters to the problem in a misguided attempt to be as realistic as possible. This often leads to excessive variance of parameters and overfitting when restraint is not exercised on small datasets. Fortunately, statistical inference techniques are accompanied by powerful model selection techniques that can be used to select the model that best explains the data without overfitting. Standard methods are traditionally used in phylogenetic inference: likelihood ratio tests (Wilks, 1938), Cox-Goldman test (Cox, 1962; Goldman, 1993), Akaike Information Criterion (Akaike, 1973), Bayesian Information Criterion (Schwarz, 1978) and Bayes' factors (Kass and Raftery, 1995). The model selection issue was bypassed to some extent in this thesis. Under the reversible jump method used to design the time-heterogeneous method in chapter 4, the evolutionary model is just another phylogenetic parameter and the number of composition vectors necessary to fit the data without overfitting is automatically determined during the inference process. Overfitting issues were encountered while building a model that allows for heterogeneity of the process across sites in chapter 5 but they have been resolved by using an appropriate prior model, based on the empirical analysis of several datasets.

Inferences based on real and synthetic datasets, have shown that using these new evolutionary models can result in visible, albeit limited, differences on inferred tree topology and clade support values. Nevertheless, the work presented in this thesis is expected to have more impact on our understanding of RNA sequence evolution than on the ability of phylogenetic methods to recover correct evolutionary relationships.

6.2 Phylogenomics and the future of PHASE

Over the past three years, it has become increasingly clear that some important phylogenetic questions could not be resolved with a single-gene approach alone. Although the **PHASE** software allows the user to combine multiple RNA genes, *e.g.*, rRNAs and tRNAs, when performing a phylogenetic analysis, the limited size of the datasets and the very nature of the evolutionary process might not always allow for the resolution of deep phylogenetic relationships with significant confidence. Reconstruction methods based on whole-genome datasets have consequently a definite advantage due to the sheer size of the datasets involved[1]. Using larger datasets clearly reduces the sampling error but also offers opportunities to detect and resolve inconsistencies between different genes. These inconsistencies might arise because of Horizontal Gene Transfer, *i.e.*, non tree-like evolutionary histories, or because of incorrect models of sequence evolution.

Obviously, using **PHASE** and using complex RNA models does not preclude us from a concatenated sequences approach with combined substitution models. If the currently implemented substitution processes are judged unsatisfactory, it is certainly possible to add better methods and models that would target the specificities of the evolution of protein-coding genes to the software. Moreover, the new models proposed here for RNA genes could certainly be adapted to improve phylogenetic methods based on protein-coding sequence data. However, dramatic improvements to heuristic tree search under ML have been seen in the recent years (Guindon and Gascuel, 2003), and one has to concede that Bayesian methods, while still very attractive for small datasets, may be slower when studying large concatenated datasets.

[1] Nevertheless, it has recently been suggested that careless gene concatenation sometimes replace weak incongruences with statistically significant ones (Jeffroy et al., 2006).

In this thesis, the focus was on the evolution of tRNA and rRNA genes, the well-studied genes that are central to the functioning of the translational apparatus. Other non-coding RNA genes have attracted lots of attention recently (Ruvkun, 2001), and we are aware of at least three independent research groups who have started using **PHASE** to study the evolution of miRNA genes recently. Obviously, these works are not aimed at improving our knowledge of systematic relationships between species but at understanding how these biologically important genes are evolving. Micro-RNAs are unfortunately quite small and consequently difficult to analyze with **PHASE**'s complex RNA models but hopefully this should not prevent interesting results from coming out.

6.3 Possible extensions and future work

There is scope for future work in the presented area of research. New evolutionary models and methods have been introduced and demonstrated on simple datasets and it would now be useful to apply them on other challenging phylogenetic problems. How ubiquitous are the trends of compositional variation across sites encountered while studying mitochondrial mammalian RNA genes? Can time-heterogeneous models recover the traces of ancient evolutionary forces?

6.3.1 Improving the MCMC algorithms

At the moment, little research has been done to guide in the choice of good proposal distributions. This is not an issue from a theoretical point of view because results do not depend on these choices with enough running time. However, the matter has probably more importance in practice and a huge amount of computational power might be wasted on inefficient proposals. It should be noted that, of all the optimizations that have been implemented in **PHASE** in the past three years, none had as much effect as improving the tuning parameters that control the proposal of new states during a MCMC run. A related problem, which was not examined in this thesis, is how to diagnose convergence of a chain. Here, the results of multiple chains were always compared for congruence but this is a rather expensive method.

Finally, one could investigate the potential of hybrid Monte Carlo methods in Bayesian phylogenetic inference (Duane et al., 1987). In chapter 3, their use was suggested for estimating branch length parameters because the mixing is often problematic with these parameters and because their partial derivatives can be computed analytically. Nevertheless, hybrid methods might also show an edge over the standard Metropolis-Hastings algorithm with the other evolutionary parameters.

6.3.2 Evolutionary models

In this thesis, **PHASE** is described as a package to perform phylogenetic inference with RNA genes because base-pair models that account for the specificity of nucleotide interactions in RNA helices have been implemented. Nevertheless, different evolutionary models can be, and are, plugged into the software. As mentioned before, implementing new models that are adapted for the evolution of protein-coding genes can often be done easily and could be useful to handle more and larger datasets.

In chapter 5, compositional heterogeneity across sites has been shown to affect the performance of time-heterogeneous methods. Combining the time-heterogeneous model with a model of compositional variation across sites seems therefore like a natural thing to do. A computationally tractable solution would first involve better MCMC proposals to solve the mixing issue encountered with the frequency parameters when a Gaussian process model is used (chapter 5).

In this thesis, the discrete gamma model was used to model the variation of the selection pressure across sites. It has been pointed out that combining a codon model with the gamma model was not ideal for phylogenetic analysis based on protein-coding DNA sequences. Models that allows for across-sites variation of ω, the nonsynonymous/synonymous ratio are more realistic (Yang et al., 1998). Similarly, one could argue that the gamma model is not appropriate to model across-site variations between paired sites in RNA sequences because pairs under weak evolutionary constraints not only evolve faster, but also tend to break more easily. A model that would allow for variation in the selection pressure for conserved secondary structure, rather than variation of the average substitution rate, seems a more sensible alternative. The 16D model (Savill et al., 2001), originally derived from the base-pair models introduced in Muse (1995), could be an interesting starting point for a potential solution. Unlike all the substitution models introduced so far, the 16D model is not parameterized with a set of equilibrium frequencies and some exchangeability parameters but with four frequency parameters (one for each nucleotide) and some extra pairing parameters (λ and θ) that model the fact that the Watson-Crick pairs and the G:U pairs are preferred to the ten other mismatch pairs. If λ and θ were allowed to vary across sites, one could simultaneously explain the presence of slow and fast evolving pairs and why the G:C+C:G frequency is higher at slow evolving sites. If these pairing parameters were also allowed to vary over lineages, this would give a time-heterogeneous model that explains the variation of G:C frequency in different lineages as a consequence of varying selection pressure over time. Finally, since the two elements of a pair behave like independent entities for specific values of these pairing parameters, one could account for the variation of the secondary structure in different lineages by combining this model with a covarion-like model (Galtier, 2001) that allows for change of the substitution process at a site over evolutionary time.

Appendix A

Jacobian for the split/merge proposals of composition vectors

We recall that the Jacobian for the transformation is (4.10):

$$J = \left| \frac{\partial({\pi_1}^{(1)}, \ldots, {\pi_{n-1}}^{(1)}, {\pi_1}^{(2)}, \ldots, {\pi_{n-1}}^{(2)}, s_1, s_2)}{\partial({\pi_1}^{(0)}, \ldots, {\pi_{n-1}}^{(0)}, u_1, \ldots, u_n, s_0)} \right| \quad .$$

This can be decomposed into:

$$J = \left| \frac{\partial({\pi_1}^{(1)}, \ldots, {\pi_{n-1}}^{(1)}, s_1, {\pi_1}^{(2)}, \ldots, {\pi_{n-1}}^{(2)}, s_2)}{\partial(m_1^{(1)}, \ldots, m_n^{(1)}, m_1^{(2)}, \ldots, m_n^{(2)})} \right| \times \left| \frac{\partial(m_1^{(1)}, \ldots, m_n^{(1)}, m_1^{(2)}, \ldots, m_n^{(2)})}{\partial(m_1^{(0)}, \ldots, m_n^{(0)}, u_1, \ldots, u_n)} \right| \times \left| \frac{\partial(m_1^{(0)}, \ldots, m_n^{(0)}, u_1, \ldots, u_n)}{\partial({\pi_1}^{(0)}, \ldots, {\pi_{n-1}}^{(0)}, s_0, u_1, \ldots, u_n)} \right| \quad .$$

Since these matrices are bloc-diagonal, this can be simplified further into:

$$J = \frac{1}{|A^{(1)}|\,|A^{(2)}|} \times \prod_{j=1}^{n} |B_j| \times |A^{(0)}| \quad , \tag{A.1}$$

where

$$A^{(i)} = \left| \frac{\partial(m_1^{(i)}, \dots, m_n^{(i)})}{\partial({\pi_1}^{(i)}, \dots, {\pi_{n-1}}^{(i)}, s_i)} \right| = \begin{vmatrix} s_i & 0 & 0 & \dots & 0 & {\pi_1}^{(i)} \\ 0 & s_i & 0 & \dots & 0 & {\pi_2}^{(i)} \\ 0 & 0 & s_i & \dots & 0 & {\pi_3}^{(i)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & s_i & {\pi_{n-1}}^{(i)} \\ -s_i & -s_i & -s_i & \dots & -s_i & {\pi_n}^{(i)} \end{vmatrix}$$

for $i = 0$, $i = 1$ and $i = 2$, and

$$B_j = \left| \frac{\partial(m_j^{(1)}, m_j^{(2)})}{\partial(m_j^{(0)}, u_j)} \right|$$

for all j in $1..n$.

Since

$$B_j = \begin{vmatrix} (1 - \frac{b_0' u_j}{2b_1' \sqrt{m_j^{(0)}}}) \ \exp(\frac{b_0' u_j}{b_1' \sqrt{m_j^{(0)}}}) & \frac{\sqrt{m_j^{(0)}} b_0'}{b_1'} \exp(\frac{b_0' u_j}{b_1' \sqrt{m_j^{(0)}}}) \\ (1 + \frac{b_0' u_j}{2b_2' \sqrt{m_j^{(0)}}}) \ \exp(-\frac{b_0' u_j}{b_2' \sqrt{m_j^{(0)}}}) & -\frac{\sqrt{m_j^{(0)}} b_0'}{b_2'} \exp(-\frac{b_0' u_j}{b_2' \sqrt{m_j^{(0)}}}) \end{vmatrix}$$
$$= \frac{{b_0'}^2 \sqrt{m_j^{(0)}}}{b_1' b_2'} \quad exp(\frac{b_0' u_j (b_2' - b_1')}{b_1' b_2' \sqrt{m_j^{(0)}}}) \quad ,$$

we only have to demonstrate that $A^{(i)} = {s_i}^{n-1}$ to prove (4.11). We demonstrate in general that

$$E(n) = \begin{vmatrix} s_i & 0 & 0 & \dots & 0 & e_1 \\ 0 & s_i & 0 & \dots & 0 & e_2 \\ 0 & 0 & s_i & \dots & 0 & e_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & s_i & e_{n-1} \\ -s_i & -s_i & -s_i & \dots & -s_i & e_n \end{vmatrix} = {s_i}^{n-1} \sum_{j=1}^{n} e_j \qquad \forall n \geq 2 \quad ,$$

which proves that $A^{(i)} = {s_i}^{n-1} \sum_{l=1}^{n} \pi_l^{(i)} = {s_i}^{n-1}$ since frequencies sum up to 1.0.

This can be done by induction.

$$E(2) = \begin{vmatrix} s_i & e1 \\ -s_i & e2 \end{vmatrix} = s_i(e_1 + e_2)$$

proves the property for $n = 2$.

Using the expansion by minors along the first line,

$$E(n+1) = \begin{vmatrix} s_i & 0 & 0 & \dots & 0 & e_1 \\ 0 & s_i & 0 & \dots & 0 & e_2 \\ 0 & 0 & s_i & \dots & 0 & e_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & s_i & e_n \\ -s_i & -s_i & -s_i & \dots & -s_i & e_{n+1} \end{vmatrix}_{n+1} =$$

$$s_i \begin{vmatrix} s_i & 0 & 0 & \dots & 0 & e_2 \\ 0 & s_i & 0 & \dots & 0 & e_3 \\ 0 & 0 & s_i & \dots & 0 & e_4 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & s_i & e_n \\ -s_i & -s_i & -s_i & \dots & -s_i & e_{n+1} \end{vmatrix}_n + (-1)^{n+2} e_1 \begin{vmatrix} 0 & s_i & 0 & 0 & \dots & 0 \\ 0 & 0 & s_i & 0 & \dots & 0 \\ 0 & 0 & 0 & s_i & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & s_i \\ -s_i & -s_i & -s_i & -s_i & \dots & -s_i \end{vmatrix}_n$$

Using the recurrence property for the first term and expanding the second determinant further along the first column gives

$$\begin{aligned} E(n+1) =& s_i \left[{s_i}^{n-1} \sum_{l=2}^{n+1} e_l \right] + (-1)^{n+2} e_1 \left[(-1)^{n+2} s_i \left| s_i I_{n-1} \right| \right] \\ =& {s_i}^n \left(\sum_{l=2}^{n+1} e_l \right) + e_1 \left[s_i \times {s_i}^{n-1} \right] \\ =& {s_i}^n \left(\sum_{l=1}^{n+1} e_l \right) \quad , \end{aligned}$$

which completes the demonstration.

Appendix B

The normalizing factor Z used for the GP prior

In equation (5.6), a factor Z was introduced to normalize the GP prior on the collection of activation values. Since Z depends on the gamma shape parameter, the prior value changes when proposing new values for α during a MCMC run. This affects the prior ratio used during the computation of the acceptance rate. The complete expression of Z is given here.

Let $\mathbf{A}$ be the collection of activation values. In the standard case,

$$Z = \int_{\mathbf{A}} d\mathbf{A} \exp\{\sum_{i=1}^{n} -\frac{1}{2}\mathbf{a_i}^T\mathbf{C_i}^{-1}\mathbf{a_i}\} \quad .$$

Following equation (6) in Neal (1997),

$$-\log(Z) = \sum_{i=1}^{n}(-\frac{C}{2}\log(2\pi) - \frac{1}{2}\log\det\mathbf{C_i}) \quad ,$$

where C is the number of discrete gamma categories and $\mathbf{C_i}$ are the C by C covariance matrices of the GP.

However, an extra constraint was added on the activation values, and, in our case:

$$Z = \int_{\mathbf{A}} d\mathbf{A} \ \exp\{\sum_{i=1}^{n} -\frac{1}{2}\mathbf{a_i}^T\mathbf{C_i}^{-1}\mathbf{a_i}\} \prod_{c=1}^{C} \delta(\sum_{i=1}^{n} a_i(r_c))$$

Using the Fourier transform expression of the delta function,

$$Z = \int_{\mathbf{A}} d\mathbf{A}\ \exp\{-\frac{1}{2}\sum_{i=1}^{n} \mathbf{a}_\mathrm{i}{}^T \mathbf{C}_\mathrm{i}{}^{-1} \mathbf{a}_\mathrm{i}\} \prod_{c=1}^{C} \int_{-\infty}^{\infty} dy\ \exp(-2\pi i y \sum_{i=1}^{n} a_i(r_c)) \quad ,$$

which gives:

$$\begin{aligned} Z &= \int_{\mathbf{A}} d\mathbf{A} \int_{\mathbf{y}} \exp\{-\frac{1}{2}\sum_{i=1}^{n} \mathbf{a}_\mathrm{i}{}^T \mathbf{C}_\mathrm{i}{}^{-1} \mathbf{a}_\mathrm{i}\} \exp\{\sum_{c=1}^{C}(-2\pi i y_c \sum_{i=1}^{n} a_i(r_c))\}\ dy_1 \ldots dy_C \\ &= \int_{\mathbf{y}} d\mathbf{y} \prod_{i=1}^{n} \int_{\mathbf{a}_\mathrm{i}} d\mathbf{a}_\mathrm{i}\ \exp\{-\frac{1}{2}\mathbf{a}_\mathrm{i}{}^T \mathbf{C}_\mathrm{i}{}^{-1} \mathbf{a}_\mathrm{i} - 2\pi i \mathbf{y}^T \mathbf{a}_\mathrm{i}\} \quad . \end{aligned}$$

We note that:

$$\begin{aligned} &-\frac{1}{2}\left(\mathbf{a}_\mathrm{i} + 2\pi i \mathbf{C}_\mathrm{i}\mathbf{y}\right)^T \mathbf{C}_\mathrm{i}{}^{-1} \left(\mathbf{a}_\mathrm{i} + 2\pi i \mathbf{C}_\mathrm{i}\mathbf{y}\right) - 2\pi^2 \mathbf{y}^T \mathbf{C}_\mathrm{i}\mathbf{y} \\ &\qquad = -\frac{1}{2}\left\{\mathbf{a}_\mathrm{i}{}^T \mathbf{C}_\mathrm{i}{}^{-1}\mathbf{a}_\mathrm{i} + 2\pi i \mathbf{y}^T \mathbf{a}_\mathrm{i} + 2\pi i \mathbf{a}_\mathrm{i}{}^T \mathbf{y} - 4\pi^2 \mathbf{y}^T \mathbf{C}_\mathrm{i}\mathbf{y}\right\} - 2\pi^2 \mathbf{y}^T \mathbf{C}_\mathrm{i}\mathbf{y} \\ &\qquad\qquad = -\frac{1}{2}\mathbf{a}_\mathrm{i}{}^T \mathbf{C}_\mathrm{i}{}^{-1}\mathbf{a}_\mathrm{i} - 2\pi i \mathbf{y}^T \mathbf{a}_\mathrm{i} \quad . \end{aligned}$$

Consequently:

$$\begin{aligned} Z &= \int_{\mathbf{y}} d\mathbf{y} \prod_{i=1}^{n} \int_{\mathbf{a}_\mathrm{i}} d\mathbf{a}_\mathrm{i}\ \exp\{-\frac{1}{2}\left(\mathbf{a}_\mathrm{i} + 2\pi i \mathbf{C}_\mathrm{i}\mathbf{y}\right)^T \mathbf{C}_\mathrm{i}{}^{-1}\left(\mathbf{a}_\mathrm{i} + 2\pi i \mathbf{C}_\mathrm{i}\mathbf{y}\right) - 2\pi^2 \mathbf{y}^T \mathbf{C}_\mathrm{i}\mathbf{y}\} \\ &= \int_{\mathbf{y}} d\mathbf{y} \prod_{i=1}^{n} \left(\exp\{-2\pi^2(\mathbf{y}^T \mathbf{C}_\mathrm{i}\mathbf{y})\} \int_{\mathbf{a}_\mathrm{i}} d\mathbf{a}_\mathrm{i}\ \exp\{-\frac{1}{2}\left(\mathbf{a}_\mathrm{i} + 2\pi i \mathbf{C}_\mathrm{i}\mathbf{y}\right)^T \mathbf{C}_\mathrm{i}{}^{-1}\left(\mathbf{a}_\mathrm{i} + 2\pi i \mathbf{C}_\mathrm{i}\mathbf{y}\right)\}\right) \quad . \end{aligned}$$

One can recognize here the characteristic function of a multivariate Gaussian random variable:

$$\begin{aligned} Z &= \int_{\mathbf{y}} d\mathbf{y} \left(\prod_{i=1}^{n} \exp\{-2\pi^2(\mathbf{y}^T\mathbf{C}_{\mathrm{i}}\mathbf{y})\}\right)\left(\prod_{i=1}^{n} \frac{(2\pi)^{C/2}}{\sqrt{\det \mathbf{C}_{\mathrm{i}}{}^{-1}}}\right) \\ &= \prod_{i=1}^{n} \frac{(2\pi)^{C/2}}{\sqrt{\det \mathbf{C}_{\mathrm{i}}{}^{-1}}} \quad \times \quad \int_{\mathbf{y}} d\mathbf{y}\ \exp\{-\frac{1}{2}(2\pi\mathbf{y})^T(\sum_{i=1}^{n}\mathbf{C}_{\mathrm{i}})(2\pi\mathbf{y})\} \\ &= \prod_{i=1}^{n} \frac{(2\pi)^{C/2}}{\sqrt{\det \mathbf{C}_{\mathrm{i}}{}^{-1}}} \quad \times \quad \frac{1}{(2\pi)^C}\frac{(2\pi)^{C/2}}{\sqrt{\det \sum_{i=1}^{n} \mathbf{C}_{\mathrm{i}}}} \\ &= (2\pi)^{\frac{C(n-1)}{2}} \times \sqrt{\frac{\prod_{i=1}^{n} \det \mathbf{C}_{\mathrm{i}}}{\det(\sum_{i=1}^{n} \mathbf{C}_{\mathrm{i}})}} \quad . \end{aligned}$$

Finally:

$$-\log(Z) = -\frac{C(n-1)}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log\det\mathbf{C}_{\mathrm{i}} + \frac{1}{2}\log\det(\sum_{i=1}^{n}\mathbf{C}_{\mathrm{i}}) \quad .$$

Bibliography

Adams, K. L. and J. D. Palmer

2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution*, 29:380–295. (Not cited.)

Akaike, H.

1973. Information theory and an extension of the Maximum Likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, Pp. 267–281. (Cited on page 149.)

Akaike, H.

1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723. (Cited on page 40.)

Alfaro, M. E., S. Zoller, and F. Lutzoni

2003. Bayes or bootstrap? a simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, 2003:255–266. (Cited on page 99.)

Baldauf, S. L. and J. D. Palmer

1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proceedings of the National Academy of Sciences*, 90:11558–11562. (Cited on page 25.)

Barry, D. and J. A. Hartigan

1987. Statistical analysis of hominoid molecular evolution. *Statistical Science*, 2:191–210. (Cited on page 79.)

Benner, S. E., C. M. A., and G. G. H.
1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Engineering*, 7:1323–1332. (Cited on page 21.)

Bergthorsson, U., K. L. Adams, B. Thomason, and J. D. Palmer
2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, 424:197–201. (Cited on page 26.)

Brooks, D. J., J. R. Fresco, and M. Singh
2004. A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. *Bioinformatics*, 20:2251–2257. (Cited on page 119.)

Bruno, W. J.
1996. Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution*, 13:1368–1375. (Cited on page 118.)

Bruno, W. J. and A. L. Halpern
1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Molecular Biology and Evolution*, 16:564–566. (Cited on page 76.)

Bruno, W. J., N. D. Socci, and A. L. Halpern
2000. Weighted Neighbor-Joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, 14:189–197. (Cited on page 29.)

Camin, J. H. and R. R. Sokal
1965. A method for deducing branching sequences in phylogeny. *Evolution*, 19:311–326. (Cited on page 27.)

Cappé, O., C. P. Robert, and T. Rydén
2003. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society Series B*, 65:679–700. (Cited on page 93.)

Carlin, B. P. and T. A. Louis
2000. *Bayes and empirical Bayes methods for data analysis*, chapman and Hall, New York edition. (Cited on page 124.)

Cavalli-Sforza, L. L. and A. W. F. Edwards

1967. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21:550–570. (Cited on pages 28 and 29.)

Chang, B. S. W. and D. L. Campbell

2000. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Molecular Biology and Evolution*, 17:1220–1231. (Cited on pages 76 and 114.)

Chang, J. T.

1996a. Full reconstruction of markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences*, 137:51–73. (Cited on page 75.)

Chang, J. T.

1996b. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences*, 134:189–215. (Cited on page 19.)

Chu, W., Z. Ghahramani, F. Falciani, and D. L. Wild

2005. Biomarker discovery in microarray gene expression data with gaussian processes. *Bioinformatics*, 21:3385–3393. (Cited on page 135.)

Conant, G. C. and P. O. Lewis

2001. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. *Molecular Biology and Evolution*, 18:1024–1033. (Cited on pages 77, 101, and 114.)

Cox, D. R.

1962. Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society Series B*, 24:406–424. (Cited on pages 40 and 149.)

Creevey, C. J., D. A. Fitzpatrick, G. K. Philip, R. J. Kinsella, M. J. O'Connell, M. M. Pentony, S. A. Travers, W. M., and J. O. McInerney

2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proceedings of the Royal Society of London B*, 271:2551–2558. (Cited on page 26.)

Crooks, G. E. and S. E. Brenner

2005. An alternative model of amino acid replacement. *Bioinformatics*, 21:975–980. (Cited on page 21.)

Darwin, C.

1859. *On the Origin of Species by Means of Natural Selection*, john Murray, London edition. (Cited on page 13.)

Di Giulio, M.

2000. The universal ancestor lived in a thermophilic or hyperthermophilic environment. *Journal of Theorical Biology*, 203:203–213. (Cited on pages 120 and 143.)

Di Giulio, M.

2003. The universal ancestor was a thermophile or a hyperthermophile: tests and further evidence. *Journal of Theorical Biology*, 221:425–436. (Cited on page 120.)

Dimmic, M. W., D. P. Mindell, and R. A. Goldstein

2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pacific Symposium on Biocomputing*, Pp. 18–29. (Cited on page 118.)

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery

2003. Comparison of Bayesian and Maximum Likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, 20:248–254. (Cited on page 99.)

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth

1987. Hybrid Monte Carlo. *Physics Letters B*, 195:216–222. (Cited on pages 57 and 152.)

Eck, R. V. and M. O. Dayhoff

1966. *Atlas of protein sequence and structure*, national Biomedical Research Foundation, Silver Spring, Maryland edition. (Cited on page 27.)

Embley, T. M., R. H. Thomas, and R. A. D. Williams

1993. Reduced thermophilic bias in the 16S rDNA sequence from Thermus ruber provides further support for a relationship between Thermus and Deinococcus. *Systematic and Applied Microbiology*, 16:25–29. (Cited on page 77.)

Erixon, P., B. Svennblad, T. Britton, and B. Oxelman

2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology*, 52:665–673. (Cited on page 99.)

Farris, J. S.

1999. Likelihood and inconsistency. *Cladistics*, 15:199–204. (Cited on page 19.)

Felsenstein, J.

1973. Maximum Likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22:240–249. (Cited on page 75.)

Felsenstein, J.

1978. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410. (Cited on pages 28 and 75.)

Felsenstein, J.

1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376. (Cited on pages 18, 45, 46, 47, 50, and 79.)

Felsenstein, J.

1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791. (Cited on page 19.)

Felsenstein, J.

2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution*, 53:447–455. (Cited on page 117.)

Felsenstein, J.

2004. *Inferring phylogenies*, sinauer Associates, Sunderland, Mass. edition. (Cited on pages 19, 27, 60, and 63.)

Felsenstein, J. and G. Churchill

1996. A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13:93–104. (Cited on page 122.)

Fitch, W. M.

1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416. (Cited on page 28.)

Fitch, W. M. and E. Margoliash

1967. Construction of phylogenetic trees. *Science*, 155:279–284. (Cited on pages 28 and 29.)

Fitch, W. M. and E. Markowitz

1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4:579–593. (Cited on page 124.)

Foster, P.

2004. Modeling compositional heterogeneity. *Systematic Biology*, 53:485–495. (Cited on pages 77, 80, 112, 113, 119, and 149.)

Foster, P. G. and D. A. Hickey

1997. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, 48:284–290. (Cited on pages 76 and 114.)

Galtier, N.

2001. Maximum Likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, 18:866–873. (Cited on page 153.)

Galtier, N.

2004. Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. *Systematic Biology*, 53:38–46. (Cited on page 38.)

Galtier, N. and M. Gouy

1995. Inferring phylogenies from DNA sequences of unequal base compositions. *Proceedings of the National Academy of Sciences*, 92:11317–11321. (Cited on pages 29, 77, and 79.)

Galtier, N. and M. Gouy

1998. Inferring pattern and process: Maximum-Likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, 15:871–879. (Cited on pages 79, 80, 105, 119, and 142.)

Galtier, N. and J. R. Lobry
1997. Relationships between genomic g+c content, rna secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, 44:632–636. (Cited on pages 104 and 105.)

Galtier, N., N. Tourasse, and M. Gouy
1999. A nonhyperthermophilic common ancestor to extant life forms. *Science*, 283:220–221. (Cited on pages 105, 106, 107, 110, 112, 114, 116, 119, 120, 141, 143, 144, and 145.)

Gardner, P. P., A. Wilm, and S. Washietl
2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, 33:2433–2439. (Cited on page 26.)

Gascuel, O.
1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14:685–695. (Cited on page 29.)

Gelman, A., G. O. Roberts, and W. R. Gilks
1996. Efficient Metropolis jumping rules. In *Bayesian Statistics*, volume 5, Pp. 599–607. Bernardo, J. and Berger, J. and Dawid, A. and Smith, A. (Cited on page 65.)

Geman, S. and D. Geman
1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741. (Cited on page 56.)

Gibbs, M. N. and D. J. C. MacKay
2000. Variational gaussian process classifiers. *IEEE-NN*, 11:1456. (Cited on page 135.)

Gibson, A., V. Gowri-Shankar, P. G. Higgs, and M. Rattray
2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. *Molecular Biology and Evolution*, 22:251–264. (Cited on pages 115, 129, and 148.)

Goldman, N.

1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, 36:182–198. (Cited on pages 19, 40, and 149.)

Goldman, N., J. L. Thorne, and D. T. Jones

1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445–458. (Cited on pages 122 and 140.)

Graur, D. and W. Martin

2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends in Genetics*, 20:80–86. (Cited on page 44.)

Green, P. J.

1995. Reversible jump Markov chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82:711–732. (Cited on pages 80, 85, and 86.)

Guindon, S. and O. Gascuel

2003. A simple, fast and accurate algorithm to estimate large phylogenies by Maximum-Likelihood. *Systematic Biology*, 52:696–704. (Cited on pages 19 and 150.)

Gutell, R. R., J. J. Cannone, Z. Shang, Y. Du, and M. J. Serra

2000. A story: unpaired adenosine bases in ribosomal rnas. *Journal of Molecular Biology*, 304:335–354. (Cited on page 128.)

Haeckel, E.

1866. *Generelle Morphologie der Organismen*, reimer, Berlin edition. (Cited on page 15.)

Halpern, A. L. and W. J. Bruno

1998. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15:910–917. (Cited on page 118.)

Harrison, G. L., P. A. McLenachan, M. J. Phillips, K. E. Slack, A. Cooper, and D. Penny

2004. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late cretaceous. *Molecular Biology and Evolution*, 21:974–983. (Cited on page 40.)

Harvey, P. H., A. J. Leigh Brown, J. Maynard Smith, and S. Nee, eds.
1996. *New Uses for New Phylogenies*, oxford University Press, Oxford edition. (Cited on page 13.)

Harvey, P. H. and M. D. Pagel
1991. *The comparative method in evolutionary biology*, oxford University Press, Oxford edition. (Cited on page 16.)

Hasegawa, M., T. Hashimoto, J. Adachi, N. Iwabe, and T. Miyata
1993. Early branchings in the evolution of Eukaryotes: ancient divergence of Entamoeba that lacks mitochondria revealed by protein sequence data. *Journal of Molecular Evolution*, 36:380–388. (Cited on page 107.)

Hasegawa, M., H. Kishino, and T. Yano
1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174. (Cited on page 34.)

Hastings, W. K.
1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109. (Cited on pages 20 and 53.)

Hedges, S. B. and S. Kumar
2003. Genomic clocks and evolutionary timescales. *Trends in Genetics*, 19(4):200–206. (Cited on page 44.)

Higgs, P. G.
1998. Compensatory neutral mutations and the evolution of RNA. *Genetica*, 102/103:91–101. (Cited on page 36.)

Hillis, D. M.
1995. Approaches for assessing phylogenetic accuracy. *Systematic Biology*, 44:3–16. (Cited on page 75.)

Huang, S. L., L. C. Wu, H. K. Laing, K. T. Pan, and J. T. Horng
2004. PGTdb: a database providing growth temperatures of prokaryotes. *Bioinformatics*, 20:276–278. (Cited on page 105.)

Hudelot, C., V. Gowri-Shankar, H. Jow, M. Rattray, and P. Higgs

2003. RNA-based phylogenetics methods: Application to mammalian mitochondrial RNA sequences. *Molecular Phylogenetics and Evolution*, 28:241–252. (Cited on pages 119, 123, 128, and 148.)

Huelsenbeck, J., B. Larget, and D. Swofford
2000. A compound Poisson process for relaxing the molecular clock. *Genetics*, 154:1879–1892. (Cited on pages 44, 95, and 102.)

Huelsenbeck, J. P.
1995. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of Maximum Likelihood over Neighbor Joining. *Molecular Biology and Evolution*, 12:843–849. (Cited on page 75.)

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro
2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution*, 21:1123–1133. (Cited on pages 34 and 80.)

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist
2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, 51:673–688. (Cited on pages 58, 72, and 99.)

Huelsenbeck, J. P. and R. Nielsen
1999. Variation in the pattern of nucleotide substitution across sites. *Journal of Molecular Evolution*, 48:86–93. (Cited on pages 118 and 132.)

Huelsenbeck, J. P. and B. Rannala
2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53:904–913. (Cited on pages 99, 100, and 101.)

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe
2006. Phylogenomics: the beginning of incongruence. *Trends in Ecology and Evolution*, 22:225–231. (Cited on page 150.)

Jermiin, L. S., S. Y. W. Ho, F. Ababneh, J. Robinson, and A. W. D. Larkum
2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology*, 53:638–643. (Cited on page 77.)

Jojic, V., N. Jojic, C. Meek, D. Geiger, A. Siepel, D. Haussler, and D. Heckerman
2004. Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics*, 20 (Suppl. 1):i161–i168. (Cited on page 46.)

Jow, H.
2003. *Bayesian phylogenetics using models of RNA evolution*. PhD thesis, Department of Computer Science, University of Manchester. (Cited on pages 21 and 148.)

Jow, H., C. Hudelot, M. Rattray, and P. G. Higgs
2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Molecular Biology and Evolution*, 19:1591–1601. (Cited on pages 38, 66, 70, 119, and 123.)

Jukes, T. H. and C. R. Cantor
1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, volume 3, Pp. 21–132. Munro, H. N., ed. (Cited on page 33.)

Kass, R. E. and A. E. Raftery
1995. Bayes factors. *Journal of the American Statistical Association*, 90:773–795. (Cited on page 149.)

Kimura, M.
1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120. (Cited on page 33.)

Kimura, M. and T. Ohta
1974. On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences*, 71:2848–2852. (Cited on page 117.)

Kishino, H. and M. Hasegawa
1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data. *Journal of Molecular Evolution*, 29:170–179. (Cited on page 19.)

Kishino, H., J. L. Thorne, and W. J. Bruno
2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution*, 18(3):352–361. (Cited on page 44.)

Kolaczkowski, B. and J. W. Thornton
2004. Performance of Maximum Parsimony and Likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431:980–984. (Cited on page 75.)

Kunin, V., L. Goldovsky, N. Darzentas, and C. A. Ouzounis
2005. The net of life: Reconstructing the microbial phylogenetic network. *Genome Research*, 15:954–959. (Cited on page 26.)

Lake, J. A.
1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences*, 91:1455–1459. (Cited on pages 29 and 79.)

Lamarck, J. B.
1809. *Philosophie Zoologique, ou Exposition des Considérations Relatives à l'Histoire Naturelle des Animaux*, dentu, Paris edition. (Cited on page 13.)

Lanave, C., G. Preparata, C. Saccone, and G. Serio
1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93. (Cited on page 33.)

Larget, B. and D. Simon
1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16:750–759. (Cited on pages 19 and 64.)

Lartillot, N. and H. Philippe
2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacment process. *Molecular Biology and Evolution*, 21:1095–1109. (Cited on pages 118 and 140.)

Lewis, P. O.
2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50:913–925. (Cited on page 25.)

Lewis, P. O., M. T. Holder, and K. E. Holsinger
2005. Polytomies and Bayesian phylogenetic inference. *Systematic Biology*, 54:241–253. (Cited on page 81.)

Li, S.

1996. *Phylogenetic tree construction using Markov chain Monte Carlo.* PhD thesis, Ohio State University, Columbus. (Cited on page 19.)

Liò, P. and N. Goldman

2002. Modeling mitochondrial protein evolution using structural information. *Journal of Molecular Evolution*, 54:519–529. (Cited on pages 20, 122, and 140.)

Lockhart, P. J., M. A. Steel, H. M. D., and D. Penny

1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, 11:605–612. (Cited on pages 29, 79, and 115.)

Loomis, W. F. and D. W. Smith

1990. Molecular phylogeny of Dictyostelium discodeum by protein sequence comparison. *Proceedings of the National Academy of Sciences*, 87:9093–9097. (Cited on page 76.)

Mace, G. M., J. L. Gittleman, and A. Purvis

2003. Preserving the Tree of Life. *Science*, 300:1707–1709. (Cited on page 14.)

Marra, M. A., S. J. M. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, et al.

2003. The genome sequence of the SARS-associated coronavirus. *Science*, 300:1399–1404. (Cited on page 14.)

Mau, B.

1996. *Bayesian phylogenetic inference via Markov chain Monte Carlo methods.* PhD thesis, University of Wisconsin, Madison. (Cited on page 19.)

Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko

2004. Comparison of site-specific rate-inference methods for protein sequences: Empirical bayesian methods are superior. *Molecular Biology and Evolution*, 21:1781–1791. (Cited on page 130.)

Mayrose, I., A. Mitchell, and T. Pupko

2005. Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *Journal of Molecular Evolution*, 60:345–353. (Cited on page 124.)

McGuire, G., M. C. Denham, and D. J. Balding
2001. MAC5: Bayesian inference of phylogenetic trees from DNA sequences incorporating gaps. *Bioinformatics*, 17:479–480. (Cited on page 48.)

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller
1953. Equations of states calculations for fast computing machines. *Journal of Chemical Physics*, 21:1087–1091. (Cited on pages 20 and 53.)

Minin, V. N., D. K. S., F. Fang, and M. A. Suchard
2005. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21:3034–3042. (Cited on page 95.)

Muse, S. V.
1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics*, 139:1429–1439. (Cited on pages 40 and 152.)

Neal, R. M.
1997. Monte Carlo implementation of Gaussian Process models for bayesian regression and classification. Technical Report 9702, Department of Statistics, University of Toronto. (Cited on page 157.)

Nielsen, R. and Z. Yang
1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–936. (Cited on page 118.)

Olsen, G. J. and C. R. Woese
1993. Ribosomal RNA: A key to phylogeny. *FASEB Journal*, 7:113–123. (Cited on page 76.)

Ou, C. Y., C. A. Ciesielski, G. Myers, C. I. Bandea, C. C. Luo, B. T. Korber, J. I. Mullins, G. Schochetman, R. L. Berkelman, A. N. Economou, et al.
1992. Molecular epidemiology of hiv transmission in a dental practice. *Science*, 256:1165–1171. (Cited on page 14.)

Pagel, M. and A. Meade
2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53:571–581. (Cited on pages 50, 118, and 140.)

Phillips, M. J., F. Delsuc, and D. Penny

2004. Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, 21:1455–1458. (Cited on page 115.)

Posada, D. and K. A. Crandall

2001. Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, 50:580–601. (Cited on page 40.)

Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa

2002. Combining multiple data sets in a likelihood analysis: which models are the best? *Molecular Biology and Evolution*, 19:2294–2307. (Cited on pages 49 and 50.)

Rasmussen, C. E. and C. K. I. Williams

2006. *Gaussian Processes for Machine Learning*, MIT press, Cambridge, Mass. edition. (Cited on page 135.)

Reeves, J. H.

1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution*, 35:17–31. (Cited on page 118.)

Reyes, A., C. Gissi, G. Pesole, and C. Saccone

1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Molecular Biology and Evolution*, 15:957–966. (Cited on page 129.)

Richardson, S. and P. J. Green

1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society Series B*, 59:731–792. (Cited on page 96.)

Rivera, M. C. and J. A. Lake

2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431:152–155. (Cited on page 26.)

Rogers, J. S.

1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Systematic Biology*, 46:354–357. (Cited on page 19.)

Ronquist, F. and J. P. Huelsenbeck

2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574. (Cited on page 21.)

Rosenberg, M. S. and S. Kumar

2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Molecular Biology and Evolution*, 20:610–621. (Cited on page 77.)

Ruvkun, G.

2001. Glimpses of a tiny rna world. *Science*, 294:797–799. (Cited on page 151.)

Rzhetsky, A.

1995. Estimating substitution rates in ribosomal RNA genes. *Genetics*, 141:771–783. (Cited on page 40.)

Rzhetsky, A. and M. Nei

1992. Statistical properties of the ordinary least-squares, generalized least-squares and minimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, 35:367–375. (Cited on page 29.)

Saitou, N. and M. Nei

1987. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425. (Cited on page 29.)

Sanderson, M. J.

1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, 14(12):1218–1232. (Cited on page 44.)

Savill, N. J., D. C. Hoyle, and P. G. Higgs

2001. RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum likelihood methods. *Genetics*, 157:399–411. (Cited on pages 39, 40, and 152.)

Schöniger, M. and A. von Haeseler

1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution*, 3:240–247. (Cited on page 40.)

Schwartzman, D. W. and C. H. Lineweaver

2004. The hyperthermophilic origin of life revisited. *Biochemical Society Transactions*, 32:168–171. (Cited on page 120.)

Schwarz, G.

1978. Estimating the dimension of a model. *Annals of Mathematical Statistics*, 6:461–464. (Cited on page 149.)

Seo, T., H. Kishino, and J. L. Thorne

2005. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proceedings of the National Academy of Sciences*, 102:4436–4441. (Cited on page 49.)

Sharp, P. M., D. L. Robertson, and H. B. H.

1995. Cross-species transmission and recombination of 'AIDS' viruses. *Philosophical transactions of the Royal Society of London B*, 349:41–47. (Cited on page 14.)

Shimodaira, H. and M. Hasegawa

1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16:1114–1116. (Cited on page 19.)

Siepel, A. and D. Haussler

2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, 21:468–488. (Cited on page 46.)

Simon, D. and B. Larget

2001. *Bayesian analysis in molecular biology and evolution (BAMBE), 2.03 beta*, department of Mathematics and Computer Science, Duquesne University edition. http://www.mathcs.duq.edu/larget/bambe.html. (Cited on page 21.)

Smith, A., T. W. H. Lui, and E. R. M. Tillier

2004. Empirical substitution models for ribosomal RNA. *Molecular Biology and Evolution*, 21:419–427. (Cited on page 48.)

Sokal, R. and C. Michener

1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438. (Cited on page 29.)

Soyer, O., M. W. Dimmic, R. R. Neubig, and R. A. Goldstein
2002. Using evolutionary methods to study g-protein coupled receptors. In *Pacific Symposium on Biocomputing*, Pp. 625–636. Altman, R. B. and Dunker A. K. and Hunter, L. and Lauderdale, K. and Klein, T. E. eds. (Cited on page 140.)

Steel, M.
2005. Should phylogenetic models be trying to "fit an elephant"? *Trends in Genetics*, 21:307–309. (Cited on page 80.)

Steel, M. A., L. A. Székely, and M. D. Hendy
1994. Reconstructing trees when sequence sites evolve at variable rates. *Journal of Computational Biology*, 1:153–163. (Cited on page 75.)

Stephan, W.
1996. The rate of compensatory evolution. *Genetics*, 144:419–426. (Cited on page 36.)

Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer
2003. Inferring spatial phylogenetic variation along nucleotide sequences: a multiple changepoint model. *Journal of the American Statistical Association*, 98:427–437. (Cited on page 80.)

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer
2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*, 18:1001–1013. (Cited on page 85.)

Sullivan, J. and D. L. Swofford
2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Systematic Biology*, 50:273–729. (Cited on pages 75 and 76.)

Suzuki, Y., G. V. Glazko, and M. Nei
2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, 99:16138–16143. (Cited on page 99.)

Swofford, D. L.
2003. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.*, sinauer Associates, Sunderland, Mass. edition. (Cited on page 28.)

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis
1996. Phylogenetic inference. In *Molecular Systematics (2nd edition)*, Pp. 407–515. Hillis, D. M. and Moritz, C. and Mable, B. K. (Cited on page 31.)

Tajima, F.
1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, 135(2):599–607. (Cited on page 44.)

Tamura, K.
1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, 9:678–687. (Cited on pages 79 and 142.)

Tamura, K. and S. Kumar
2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Molecular Biology and Evolution*, 19:1727–1736. (Cited on page 79.)

Tamura, K. and M. Nei
1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10:512–526. (Cited on page 34.)

Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala
2000. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the Drosophila saltans and willistoni groups, a case study. *Molecular Phylogenetics and Evolution*, 16:344–349. (Cited on page 76.)

Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala
2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. *Molecular Biology and Evolution*, 18:1464–1473. (Cited on page 114.)

Taylor, D. J. and W. H. Piel
2004. An assessment of accuracy, error, and conflict with support values from

genome-scale phylogenetic data. *Molecular Biology and Evolution*, 21:1534–1537. (Cited on page 99.)

Thorne, J. L., N. Goldman, and D. T. Jones
1996. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673. (Cited on pages 20, 122, and 140.)

Thorne, J. L., H. Kishino, and I. S. Painter
1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12):1647–1657. (Cited on page 44.)

Tillier, E. R. M.
1994. Maximum likelihood with multiparameter models of substitution. *Journal of Molecular Evolution*, 39:409–417. (Cited on pages 39 and 40.)

Tillier, E. R. M. and R. A. Collins
1995. Neighbor Joining and Maximum Likelihood with RNA sequences. *Molecular Biology and Evolution*, 12:7–15. (Cited on page 38.)

Tillier, E. R. M. and R. A. Collins
1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics*, 148:1993–2002. (Cited on pages 39, 40, 41, and 42.)

Wallace, I. M., G. Blackshields, and D. G. Higgins
2005. Multiple sequence alignments. *Current Opinion in Structural Biology*, 15:261–266. (Cited on page 26.)

Wang, H. and D. A. Hickey
2002. Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. *Nucleic Acids Research*, 30:2501–2507. (Cited on page 110.)

Watson, J. D. and F. H. C. Crick
1953. Molecular structure of nucleic acids. *Nature*, 171:737–738. (Cited on page 35.)

Whelan, S. and N. Goldman
1999. Distributions of statistics used for the comparison of models of sequence

evolution in phylogenetics. *Molecular Biology and Evolution*, 16:1292–1299. (Cited on page 123.)

Wilks, S. S.
1938. The large sample distribution of the likelihood ratio for testing composite hypothesis. *Annals of Mathematical Statistics*, 9:60–62. (Cited on page 149.)

Williams, C. K. I. and C. E. Rasmussen
1996. Gaussian processes for regression. In *Advances in Neural Information Processing Systems*, volume 8. Touretzky, D. S. and Mozer, M. C. and Hasselmo, M. E., eds. (Cited on page 135.)

Woese, C. R.
2000. Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences*, 97:8392–8396. (Cited on page 26.)

Woese, C. R., L. Achenbach, P. Rouviere, and L. Mandelco
1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. *Systematic and Applied Microbiology*, 14:364–371. (Cited on page 115.)

Woese, C. R. and G. E. Fox
1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74:5088–5090. (Cited on page 16.)

Wuyts, J., G. Perrière, and Y. Van de Peer
2004. The european ribosomal rna database. *Nucleic Acids Research*, 32:D101–D103. (Cited on pages 102 and 104.)

Yang, Z.
1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, Pp. 1396–1401. (Cited on page 121.)

Yang, Z.
1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314. (Cited on pages 106, 118, 119, and 122.)

Yang, Z.
1996a. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, 11:367–372. (Cited on page 117.)

Yang, Z.
1996b. Maximum likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution*, 42:587–596. (Cited on pages 49 and 50.)

Yang, Z.
1997a. How often do wrong models produce better phylogenies? *Molecular Biology and Evolution*, 14:105–108. (Cited on pages 19 and 76.)

Yang, Z.
1997b. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences*, 13:555–556. (Cited on page 72.)

Yang, Z.
2001. Maximum likelihood analysis of adaptive evolution in HIV-1 GP120 gene. In *Pacific Symposium on Biocomputing*, Pp. 226–37. (Cited on page 14.)

Yang, Z.
2005. Bayesian inference in molecular phylogenetics. In *Mathematics of Evolution and Phylogeny*, Pp. 63–90. Gascuel, O. (Cited on page 63.)

Yang, Z., N. Goldman, and A. Friday
1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution*, 11:316–324. (Cited on page 132.)

Yang, Z., R. Nielsen, and M. Hasegawa
1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution*, 15:1600–1611. (Cited on page 152.)

Yang, Z., R. Nielsen, G. N., and A. M. Krabbe Pedersen
2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449. (Cited on page 118.)

Yang, Z. and B. Rannala
1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain

Monte Carlo method. *Molecular Biology and Evolution*, 14:717–724. (Cited on pages 19 and 58.)

Yang, Z. and B. Rannala
2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology*, 54:455–470. (Cited on pages 60 and 100.)

Yang, Z. and D. Roberts
1995. On the use of nucleic acid sequences to infer early branches in the tree of life. *Molecular Biology and Evolution*, 12:451–458. (Cited on pages 79, 80, and 119.)

Yang, Z. and T. Wang
1995. Mixed model analysis of DNA sequence evolution. *Biometrics*, 51:552–561. (Cited on page 124.)

Yap, V. B. and T. Speed
2005. Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evolutionary Biology*, 5:2. (Cited on page 80.)

Zwickl, D. J. and M. T. Holder
2004. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Systematic Biology*, 53:877–888. (Cited on pages 61, 62, and 100.)