THE UNIVERSITY
of EDINBURGH

Statistical Approaches to Viral Phylodynamics

Luiz Max Fagundes de Carvalho

THE UNIVERSITY
of EDINBURGH

Thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy
to the
University of Edinburgh — 2018

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Luiz Max Fagundes de Carvalho

4th December 2018

In Chapter 2, the idea to show that the mapping from phylogenies to inter-coalescent intervals and numbers of lineages is surjective non-injective was suggested to me by Professor Marc Suchard (UCLA).

Chapter 4 has been published as part of Dudas, G., Carvalho, L. M., Bedford, T. et al.(2017) *Nature*, 544(7650):309–315.

Chapter 5 has been published as part of Diehl, W., Lin, A.E., Grubaugh, N.D., Carvalho, L.M. et al. (2016) *Cell*, 167(4):1088–1098.

During the PhD I was a co-author in Rambaut, A., Lam, T. T., Carvalho, L. M., Pybus, O. G. (2016) *Virus Evolution* 2: vew007 and Dudas, G., Carvalho, L. M., Rambaut, A., Bedford, T. (2018) *ELife*, 7, e31257, which are not covered in this thesis.

Abstract

The recent years have witnessed a rapid increase in the quantity and quality of genomic data collected from human and animal pathogens, viruses in particular. When coupled with mathematical and statistical models, these data allow us to combine evolutionary theory and epidemiology to understand pathogen dynamics. While these developments led to important epidemiological questions being tackled, it also exposed the need for improved analytical methods. In this thesis I employ modern statistical techniques to address two pressing issues in phylodynamics: (i) computational tools for Bayesian phylogenetics and (ii) data integration. I detail the development and testing of new transition kernels for Markov Chain Monte Carlo (MCMC) for time-calibrated phylogenetics in Chapter 2 and show that an adaptive kernel leads to improved MCMC performance in terms of mixing for a range of data sets, in particular for a challenging Ebola virus phylogeny with 1610 taxa/sequences. As a trade-off, I also found that the new adaptive kernels have longer warm up times in general, suggesting room for improvement. Chapter 3 shows how to apply state-of-the-art techniques to visualise and analyse phylogenetic space and MCMC for time-calibrated phylogenies, which are crucial to the viral phylodynamics analysis pipeline. I describe a pipeline for a typical phylodynamic analysis which includes convergence diagnostics for continuous parameters and in phylogenetic space, extending existing methods to deal with large time-calibrated phylogenies. In addition I investigate different representations of phylogenetic space through multi-dimensional scaling (MDS) or univariate distributions of distances

to a focal tree and show that even for the simplest toy examples phylogenetic space remains complex and in particular not all metrics lead to desirable or useful representations. On the data integration front, Chapters 4 and 5 detail the use data from the 2013-2016 Ebola virus disease (EVD) epidemic in West Africa to show how one can combine phylogenetic and epidemiological data to tackle epidemiological questions. I explore the determinants of the Ebola epidemic in Chapter 4 through a generalised linear model framework coupled with Bayesian stochastic search variable selection (BSSVS) to assess the relative importance climatic and socio-economic variables on EVD number of cases. In Chapter 5 I tackle the question of whether a particular glycoprotein mutation could lead to increased human mortality from EVD. I show that a principled analysis of the available data that accounts for several sources of uncertainty as well as shared ancestry between samples does not allow us to ascertain the presence of such effect of a viral mutation on mortality. Chapter 6 attempts to bring the findings of the thesis together and discuss how the field of phylodynamics, in special its methodological aspect, might move forward.

Lay Summary

Understanding the factors driving the emergence and spread of infectious diseases in an ever more globalised world is of utmost importance. In recent years, scientists have explored the information contained in the genetic material – DNA and RNA – of pathogens (HIV, Influenza, Ebola, etc) to reveal the patterns of global spread of these disease-causing entities and the factors driving their emergence (climate, human behaviour, adaptation to new hosts, etc). Making sense of all of the data, however, involves a lot of maths and computer science, and methodological innovation is being outpaced by the growth of data – in both size and quality. The research in this thesis aims at bridging that gap between the data we have and the questions we would like to answer through the development of new statistical techniques to combine data and models. I develop more efficient methods for (re)constructing the ancestry between organisms (*phylogenetic trees*) which is an essential tool in the analysis of genomic data. I also explore new ways of visualising data from computational analyses in order to aid scientists determine when they can trust their results. Finally, I use modern statistical techniques to answer two important epidemiological questions about the 2013-2016 Ebola virus disease (EVD) epidemic in West Africa, the largest in history so far. First, I investigate the factors that contributed to the epidemic and find that some regions that report no EVD cases were predicted to have high epidemic potential connected mostly with climatic factors such as seasonal temperature variation and rain and socio-economic factors such as the distance to large settlements. The lack of overlap between areas with high predicted numbers

of cases and persistence of the virus can explain why the epidemic did not spread further. I then ask: did Ebola adapt to kill humans? While the answer is probably not, my findings teach us valuable lessons about the need to properly accommodate uncertainty in observational studies in order to make valid scientific statements. Overall, the findings in this thesis demonstrate the potential of statistical methods in aiding epidemiological inference while also highlighting just how much we still need to learn.

Acknowledgements

Many people have, in one way or another, contributed for me to be in a position to do the research I describe in this thesis. From the servitor to the cleaning staff to numerous fellow scientists that during millennia pursued Truth, and on whose mighty shoulders I have stood. Andrew Rambaut is one of these rare people that possess both a powerful intellect and a kind soul. For his generosity in dedicating countless hours to discussing various aspects of my research and helping me grow as scientist, I'm thankful. Special thanks are in order to Gytis Dudas, Darren Obbard, Jarrod Hadfield, Richard Whittet, Tom Booker, Trevor Bedford, Matthew Hall, Guy Baele, Philippe Lemey and Marc Suchard for stimulating discussions. Thanks to Lisa Valentina Gecchele for single-handedly handing in this thesis.

I could acknowledge a bunch of my plonker friends by name, but this would inevitably lead to me forgetting someone(s) and getting in trouble. So I won't. You know who you are. My loving wife's support was vital over these arduous years. To you, *Tatu*, I'm very very grateful. The unwavering support of my parents and siblings not only during my PhD but throughout my whole life helped me realise my boyhood dream: becoming a professional scientist.

The quotes at the beginning of each chapter are loosely related with the topic of the chapter in question, but don't waste your time trying to make a connection if one is not obvious to you; chances are it only makes sense to myself.

Dedication

I dedicate this thesis to my parents: Esmeralda, Judy and José.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Aspetti, signorina,
le dirò con due parole
chi son, e che faccio,
come vivo. Vuole?

Rodolfo introduces himself to Mimi in the first act of *La Boheme* by Giacomo Puccini (1858–1924).

In this chapter I provide a description of the key concepts related to the research tackled in this thesis, in addition to the necessary technical background. While I touch on the overall motivation for my research along the way (specially in Section 1.6), the specific characterisation of the problems investigated in the thesis is left to the Introduction/Background section in each chapter.

1.1 Viral phylodynamics

RNA viruses (e.g. HIV, Influenza, MERS-CoV, Ebola virus) are amongst the leading causes of morbidity and mortality in humans and livestock (Woolhouse, 2002). The

combination of high mutation rates and low generation times means that these pathogens evolve at a time scale such that their genomes can be used to detect the effects of epidemiological and ecological events (Drummond et al., 2003; Grenfell et al., 2004; Duffy et al., 2008; Pybus and Rambaut, 2009). The recent years have witnessed an unprecedented increase in the availability of molecular sequences from thousands of organisms, specially from fast-evolving RNA viruses (Benson et al., 2014). This growth in the availability and quality of data has in turn made it possible to expand the repertoire of scientific questions that can be asked: what factors drive virus emergence within and transmission between populations (Dudas et al., 2017, 2018)? How do population dynamics shape viral circulation patterns (Volz et al., 2013; Bedford et al., 2015)?

To tackle these questions it is necessary to bridge the fields of evolutionary biology and epidemiology on what is now known as "viral phylodynamics" (Grenfell et al., 2004; Volz et al., 2013; Pybus et al., 2013). One commonly accepted definition of phylodynamics states that it is concerned with the "study of how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies" (Grenfell et al., 2004). As a theoretical framework, phylodynamics couples phylogeny-generating models (e.g. coalescent, birth-death, etc) and mathematical modelling to understand how population or epidemic dynamics map onto phylogenies, and how to incorporate data from several sources into a coherent inference framework (Kühnert et al., 2011). However, as argued by Hall (2015) (Chapter 1, page 2), phylodynamics can also be understood as employing phylogenetic methods to obtain estimates of the ancestry between pathogen isolates and the timing of the lineage-splitting events, and then using this information to inform epidemiological inference, an approach called "phylogenetic epidemiology" by Kühnert et al. (2011). Another instance of phylogenetic models being used to inform epidemiological inference without reference to a unified theoretical framework is the estimation of past population dynamics using coalescent methods (see Section 1.3.2) which are then compared with epidemiological time series

in order to gain epidemiological insight (see Figure 3 in Bennett et al. (2009) for an example).

Figure 1.1 (Suchard et al., 2018) showcases the epidemiological analysis of 1610 Ebola virus (EBOV) complete genomes originally carried out by (Dudas et al., 2017). Using a generalised linear model (GLM) framework in conjunction with a phylogeny estimated from the genetic data, it is possible to reconstruct the spatial dynamics of EBOV as well as to study the association of several epidemiological predictors with viral spatial spread (bottom left panel). Estimating a time-calibrated phylogeny (see below) also allows one to reconstruct the past population dynamics of the virus.

The chief goal of **dating methods** is to combine information contained in the genetic divergence between isolates with some source of external timing information in order to reconstruct time-calibrated phylogenies and, in turn, infer node ages (divergence times) and evolutionary rates. Common sources of calibration information are the sampling dates in serially-sampled data sets and bounds/distributions for specific divergence events (e.g. the divergence between humans and chimpanzees), commonly used in studies where fossil information is available (Ho and Phillips, 2009). A central object in phylodynamic investigations is the **time-calibrated phylogeny**, a rooted phylogenetic tree in which the branch lengths are measured in units of calendar time (see Section 1.3.1 below for a mathematical description). These objects are specially useful because they allow one to estimate the timing of epidemiologically relevant events and gain insight into the *tempo* of epidemics.

All dating methods rely on the assumption of a "molecular clock", that is, the assumption that mutations accumulate steadily at a particular rate through time (Welch and Bromham, 2005)[1]. In their seminal work, Zuckerkandl and Pauling (1962) dated the origins of different globins by assuming a uniform – through time –

[1]Note that the molecular clock is not a metronome; mutations occur randomly rather than at deterministic intervals.

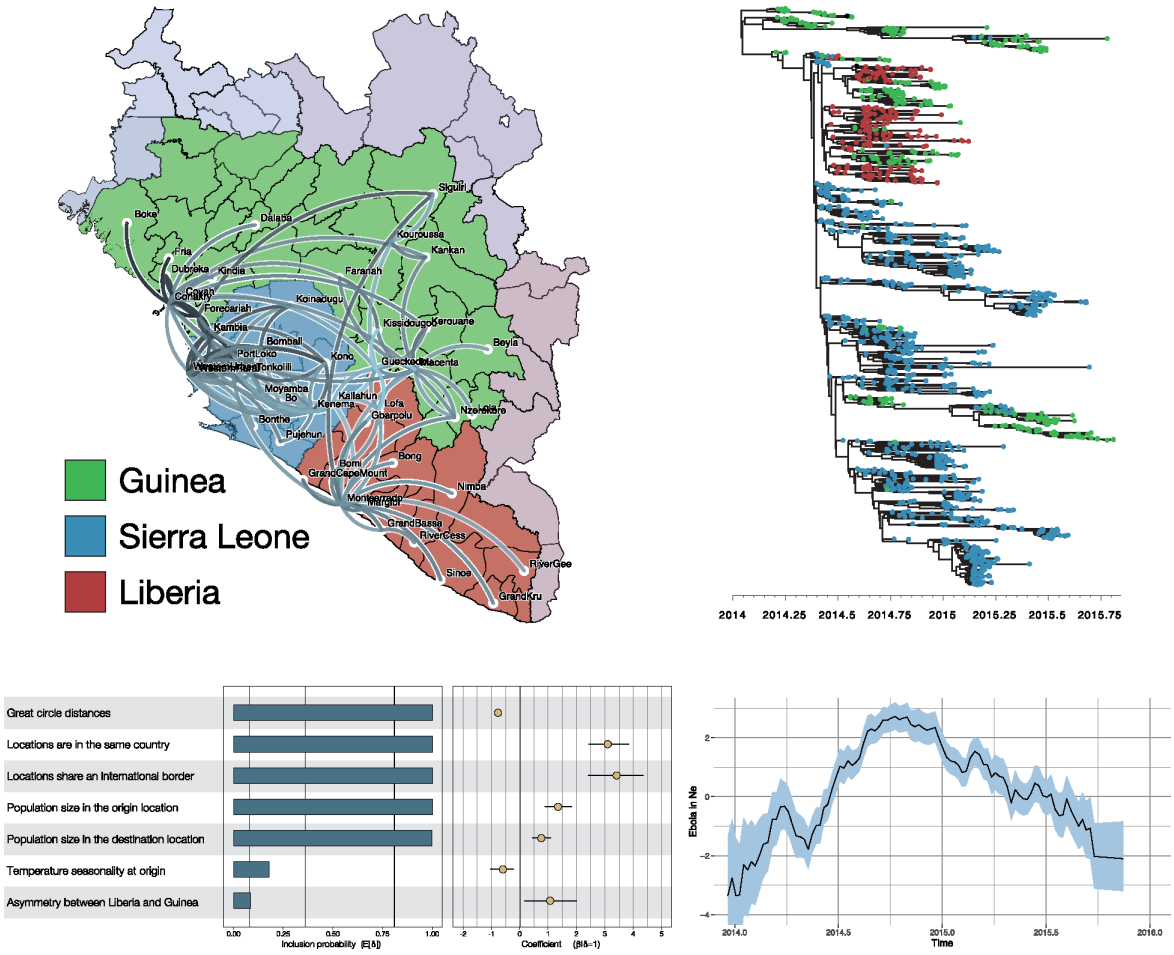


Figure 1.1: Phylodynamic analyses of the Ebola virus epidemic in West Africa with BEAST. Modern phylodynamic methods can be used to obtain insight into the spatial dynamics (top left panel), ancestry (top right, see also Figure 1.2), epidemiological determinants (bottom left) and population dynamics (bottom right) of pathogens. Here I show the analysis of 1610 complete Ebola virus genomes sampled in West Africa. Reproduced with permission from Suchard et al. (2018); please see original publication for details.

rate of molecular evolution among species and duplicated genes. This basic assumption implies that one can measure the time t of divergence between any two taxa as a function of the number of differences d between their sequences, using a rate μ to convert between the scales of expected mutations and time. There are a plethora

of dating methods, ranging from maximum likelihood (ML) (Rambaut, 2000; Volz and Frost, 2017; Sagulenko et al., 2018) to Bayesian approaches (Drummond et al., 2006) – see Table 1 in Welch and Bromham (2005) and Ho and Duchêne (2014) for surveys. A substantial amount of work has also been done on fast, approximate, regression-based approaches methods that allow treatment of large data sets (thousands of tips/samples) in reasonable time (To et al., 2015; Rambaut et al., 2016). These methods employ least squares to fit a (possibly modified) linear regression model:

$$E[d_i] = \mu\,(t_i - t_\rho)\,,$$

where d_i is the divergence (distance) between node i and the root ρ, t_i is the age of the node and t_ρ is the age of the root – also called the tree's tMRCA. In this model, μ is the slope of the regression model and the evolutionary rate. However, due to the inherent stochasticity of the mutation process, it is possible for lineages isolated closer to the age of the root to have higher divergences than younger lineages, what would lead to negative estimates of μ, which are not biologically plausible. Another limitation of regression-based methods is that the dependence between isolates is a direct violation of basic assumptions in linear regression, leading to biases in the estimation of the evolutionary rate (regression slope). For a brief discussion of the pitfalls of regression-based methods to infer evolutionary rates, see To et al. (2015) and Rambaut et al. (2016).

It is worth noting that the validity of molecular clock models has been intensely debated since the early days of its proposition (Ayala, 1999). Under a neutral (or constant selection) model, the expected number of mutational differences follows a Poisson distribution, whose mean and variance coincide. The observation that the variance usually exceeds the mean in many real world data sets has fostered intense discussions on whether the neutral model upon which molecular clocks rest is appropriate[2]. In many situations, there is substantial variation in evolutionary rates

[2]As pointed out by Ho and Larson (2006), most controversies seem to stem from the failure to account for time-dependency of evolutionary rates.

between lineages, violating the assumption of a fixed, unique rate of evolution. This needs to be accounted for in order to obtain correct estimates of node ages. Several rate-smoothing methods have been developed to model the molecular clock and between-lineage variation, and models of molecular clock can be non-exhaustively divided in the following categories(Ho et al., 2015)[3]:

- Strict clock: a single rate is assumed for all the branches of a phylogeny, as discussed above (Zuckerkandl and Pauling, 1962).
- Local clock: a fixed number of strict clocks in the tree, so that the rate is constant for some lineages (Yoder and Yang, 2000; Drummond and Suchard, 2010).
- Autocorrelated models: rates vary gradually along a lineage and thus show some degree of correlation along the phylogeny (Thorne et al., 1998).
- Uncorrelated models, where the rates for each of the branches are independently and identically distributed random variables, e.g. the uncorrelated log-normal clock (Drummond et al., 2006).

Please consult Kumar (2005) for a rather enjoyable historical account of the evolution of molecular clocks over most of the last half of the 20th century.

In this thesis I shall concentrate on the uncorrelated "relaxed" clocks of Drummond et al. (2006), because of their flexibility. Consider a rooted tree with n taxa. This tree will have $N = 2n-3$ branches excluding the root, for which we will not estimate the rate. Drummond et al. (2006) assume each branch evolves according to its own rate $r_i \in \mathbf{r} = \{r_1, r_2, \ldots, r_N\}$ and that the r_i are i.i.d. random variables. The first model to consider is the **uncorrelated exponential (UCED)** rates model, where

$$r_i \sim \text{exponential}(\lambda).$$

[3]Classification adapted from `https://github.com/sebastianduchene/NELSI/blob/master/README.md`.

Let $f_{\mathbf{r}}(\cdot)$ the prior density of the rates. Then

$$f_{\mathbf{r}}(\mathbf{r}) = \prod_{i=1}^{N} \lambda e^{-\lambda r_i}, \tag{1.1}$$

$$= \lambda^N e^{-\lambda S}, \tag{1.2}$$

from which one can note that the prior density depends on the rates $\mathbf{r}$ only through their sum $S = \sum_{i=1}^{N} r_i$, which makes clear the potential for identifiability problems. See Rannala (2002) for more on the identifiability of phylogenetic models.

Next, consider a more flexible model, where the rates are drawn from a **log-normal distribution (UCLN)**:

$$r_i \sim \text{log-normal}(\mu, \sigma^2),$$

$$f'_{\mathbf{r}}(\mathbf{r}) = \prod_{i=1}^{N} \frac{1}{r_i\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln r_i - \mu)^2}{2\sigma^2}\right), \tag{1.3}$$

$$= \frac{1}{P \cdot (2\pi)^{N/2} \cdot \sigma^N} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\ln r_i - \mu)^2\right), \tag{1.4}$$

where $P = \prod_{i=1}^{N} r_i$, which appears identifiable at first glance.

These models are important in that they allow different lineages to evolve at rates that sometimes differ by orders of magnitude. In a sense they capture the biological variability in evolutionary rates that is possible within a population, specially when considering fast evolving viruses.

The conjunction of time calibration and relaxed clocks allows us to estimate the age of the common ancestor of the circulating strains, as well as identify lineages that evolve faster than others. In an epidemiological context, such information may prove useful when designing strategies to mitigate disease transmission, for instance. Figure 1.2 is meant to illustrate one of the end products of a phylodynamic analysis using BEAST (Drummond and Rambaut, 2007; Drummond et al., 2012), with a

time-calibrated phylogeny built using sequences from a fast-evolving RNA virus (Dengue virus serotype 2) and using a relaxed clock (log-normal) model. This phylogeny was estimated from a MCMC sample of the posterior of phylogenies, meaning it carries the uncertainty inherently associated with not knowing the true phylogenetic relationship between strains, *i.e.*, phylogenetic uncertainty (see more below).

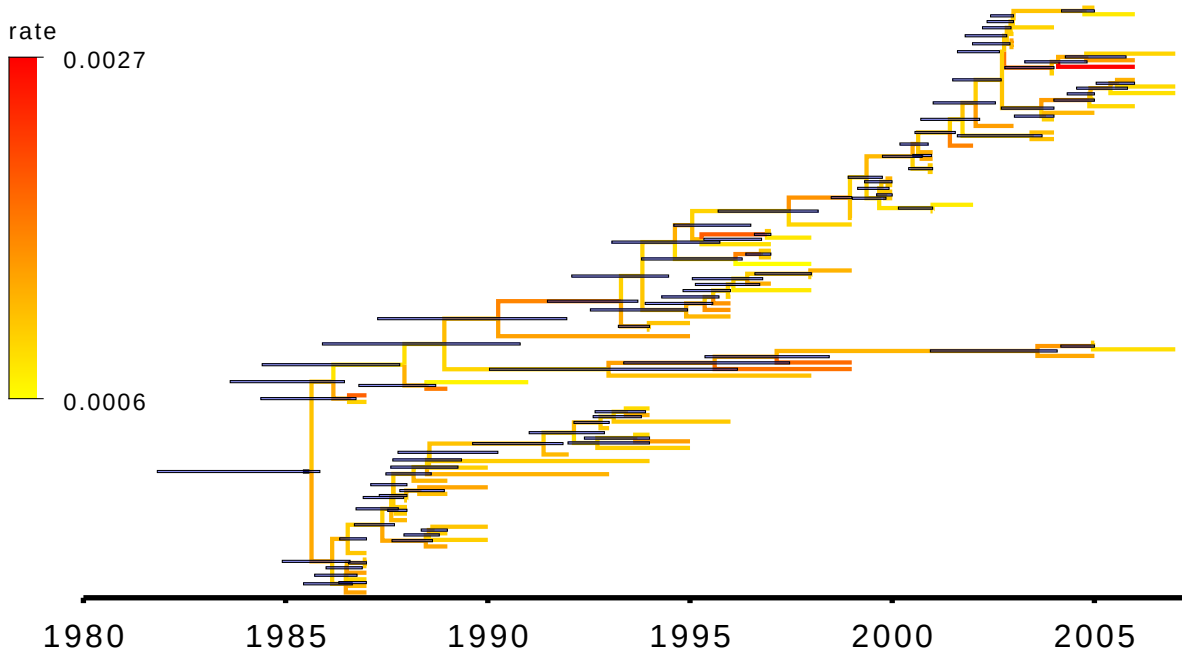


Figure 1.2: Time-calibrated phylogeny of DENV-2 strains circulating in the Americas. Phylogeny constructed using 90 *env* gene sequences from dengue virus serotype 2 (DENV-2) strains isolated in the Americas over a large time span. Branches are coloured according to their evolutionary rates (posterior mean) and the blue horizontal bars are the 95% highest probability density intervals for the node ages.

The growing mass of available data calls for the development of more realistic evolutionary models that can in turn be used to improve phylodynamic inference, while being statistically principled and computationally tractable (Pybus et al., 2013). Frost et al. (2015) list eight challenges in phylodynamic inference, of which one is of special importance to the work presented here: "how can analytical

approaches keep up with advances in sequencing?" Moreover, recent efforts have also advocated for the real-time analysis of sequence data as way of obtaining early insight into epidemic dynamics and inform intervention (Quick et al., 2016; Dudas et al., 2017; Hadfield et al., 2017). The combination of rapid increase in data and the necessity of timely analyses leads to an increased demand for more efficient estimation methods.

A key assumption for many phylodynamic methods is that the phylogeny is a good *proxy* for the **dependence structure** of the data, even if it is rarely framed in this way. A good example of this statistical interplay is given in Pybus et al. (2012), where the authors use time-calibrated phylogenies together with the sequences sampling locations and to obtain estimates of spatial epidemiological parameters such as the velocity of the epidemic wave front and the (spatial) diffusion coefficient for the West Nile virus (WNV) epidemic in the United States of America. Pybus et al. (2012) employ the phylogeny estimated from full WNV genomes as a proxy for the latent (unobserved) spatial dependence between cases and then proceed to **condition** on the phylogeny in order to infer the epidemiological parameters. Similarly, methods such as the continuous-time Markov chain (CTMC) phylogeography approach of Lemey et al. (2009) rely on the estimated phylogeny – and traits observed at the tips/leaves – to estimate migration (transition) rates between locations. This is a key concept in phylodynamics: even if the phylogeny is but a nuisance parameter, one needs to ensure it is estimated correctly – and uncertainty properly accounted for – so that inferences made conditional on the underlying dependence structure are correct. Consequently, even if one is interested solely in epidemiological parameters, properly accommodating uncertainty about the phylogeny *via* marginalisation is crucial to principled modelling. As I will show in the following sections, Bayesian methods provide a way of accounting for uncertainty about parameters and models and incorporate information from different sources, being particularly useful in phylogenetics and phylodynamics.

1.2 Bayesian/Laplacian approach to inference

This thesis is fundamentally a description of how modern statistical methods can be applied to the emerging field of phylodynamics to facilitate epidemiological inference. As such, it needs a detailed description of the underlying approach to statistical inference adopted. In what follows I shall present a short overview of the philosophical and historical roots of the Bayesian paradigm and then move on to present the necessary technical background.

1.2.1 A philosophical (and historical) digression

Statistics can understood under two main paradigms or schools of thought: orthodox/frequentist and Bayesian/Laplacian. The contrasts between these schools stem from fundamental disagreements on the nature and meaning of probability in modelling and interpreting reality. Under frequentism, probabilities are understood as long run frequencies of events while under (most versions of) Bayesianism, probabilities are understood as degrees of belief (Lindley, 2000). While the two schools differ in their interpretation of the meaning of probability, there is complete agreement in the *computation* of probabilities once a model has been established. Moreover, in practice – *i.e.* real problem-solving as opposed to musing over toy problems – there is substantial overlap and "hybrid" approaches often succeed (Kass, 2011).

A fair and detailed comparison of these approaches is well beyond the scope of this chapter and thesis. While preferring the Bayesian approach myself, I often find common – short – justifications of Bayesian inference to be wanting. I will instead choose to gloss over important objections and assume the Bayesian paradigm for statistical inference without further justification. I urge the reader to consult Chapter 11 of Robert (2007) for a grounded and well constructed defence of the Bayesian approach. Jaynes (2003) also offers a rather assertive argument in favour of Bayesianism. For a grounded and modern defence of frequentism, please see Mayo

and Spanos (2011) and references therein. Finally, for a modern and detailed account of *how* to do Bayesian statistics I recommend Gelman et al. (2014).

The name "Bayesian" – initially used in a derogatory manner – stems from a paper published by Thomas Bayes (1702–1761) in 1763 (Bayes and Price, 1763)[4] on so-called "inverse probability", a concept emerging directly from the definition of conditional probability. It can however be argued that Pierre Simon Laplace (1749–1827) was really the first to apply mathematically rigorous "inverse probability" models to scientific problems (see e.g. Laplace (1774)[5]). It is therefore my humble opinion that we owe as much or more to Laplace as we do to Reverend Bayes for laying out the foundations of (Bayesian) statistical thinking. Thus a perhaps more accurate name for this approach would be Bayesian-Laplacian. For convenience, however, I shall refer to this approach as *Bayesian* from here onwards.

1.2.2 Bayes' rule

Suppose one has a (probabilistic) model that describes how a random variable $\boldsymbol{Y}$ relates to a set of parameters $\boldsymbol{\theta}$ through a **likelihood** function $f(\boldsymbol{Y}|\boldsymbol{\theta})$. Suppose further that one observes a set of data $\boldsymbol{y}$. Under the Bayesian approach, one aims to use *Bayes' rule* to construct the **posterior** distribution

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

often written as $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. This construction of the inference problem necessitates a probability measure over the parameter space, *i.e.*, a **prior**, $\pi(\boldsymbol{\theta})$. In scientific applications, a common interpretation is that $\pi(\boldsymbol{\theta})$ encodes our knowledge about the parameters $\boldsymbol{\theta}$ *before* we observe the data $\boldsymbol{y}$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ represents our updated beliefs/knowledge.

[4] The attentive reader will notice that Bayes was dead by the time the paper was published. His friend Richard Price finished the paper and read it in front of the Royal Society.

[5] An English translation is provided in Stigler (1986).

Two points are worth emphasising: firstly, in contrast to *maximum likelihood* and other approaches, Bayesian inference focuses on *integrating* over the likely values of the quantities of interest rather than *maximising*, that is, finding the value $\hat{\boldsymbol{\theta}}$ that maximises the likelihood. Secondly, all inferences are based on the posterior; point estimates can be derived as expectations[6] and any functionals of interest can be studied by transforming $p(\boldsymbol{\theta}|\boldsymbol{y})$. The computation of posterior distributions is difficult for all but the simplest models of interest in practice. Often, these distributions are not available analytically and must be approximated. In Section 1.4 I discuss a popular method for obtaining such approximations, which will be the main focus of Chapter 2.

1.3 Bayesian phylogenetics

While the application of Bayesian methods in phylogenetics is not without controversy (Barker, 2015), this approach has enjoyed notable popularity in phylogenetics. After the pioneering work of Kuhner et al. (1995), the mid 1990s to early 2000s saw a period of rapid development of Bayesian methods to solve phylogenetic problems (Rannala and Yang, 1996; Mau and Newton, 1997; Yang and Rannala, 1997; Kuhner et al., 1998; Larget and Simon, 1999; Li et al., 2000; Suchard et al., 2001; Drummond et al., 2002). See Huelsenbeck et al. (2001) and Holder and Lewis (2003) for reviews of those developments.

The main idea behind Bayesian phylogenetics is to compute a posterior distribution of the form:

$$p(T, \theta|D) \propto f(D|T, \theta)\pi(T, \theta) \tag{1.5}$$

where D is the observed data, usually an alignment of DNA sequences, T is

[6] Mathematically, expectations are integrals.

a phylogeny and θ represents parameters such as the evolutionary rate, transition/transversion rate and coalescent parameters (see below). The likelihood is usually based on continuous-time Markov chain (CTMC) models of evolution (Hasegawa et al., 1985; Tavaré, 1986) and be efficiently computed using a dynamic programming procedure called "Felsenstein's pruning algorithm" (Felsenstein, 1981). The joint prior $\pi(T, \theta)$ can be constructed in several ways, discussed in more detail below (Section 1.3.2). This joint posterior can then be used to test evolutionary hypotheses, draw inference about population dynamics and obtain estimates of quantities of direct interest, such as the rate of substitution. In many applications we are mainly interested in θ, with T being a nuisance parameter. The main methodological problem then is accommodating phylogenetic uncertainty – *i.e* uncertainty about the true underlying ancestry that generated the data – by *marginalising* over the distribution of phylogenies to obtain a distribution $p(\theta|D)$. I shall return to the point of efficient marginalisation in Section 1.4.

In addition to providing a principled framework for accommodating uncertainty, Bayesian phylogenetics also allows for the estimation of directly interpretable quantities, such as the posterior probability for a given clade, which are useful in assessing evolutionary hypotheses. As mentioned above, the focus on marginalisation naturally leads to measures of uncertainty. While authors such as Huelsenbeck et al. (2002) have stated that Bayesian methods are also more computationally efficient[7], I argue that the main advantage of the Bayesian approach – and perhaps its greatest weakness – is the ability of the researcher to incorporate substantive expert knowledge into the analysis *via* careful construction of the prior(s).

As the field progressed and Bayesian methods became the *de facto* standard, researchers began to scrutinise the construction of prior distributions and their effects on inferences (Huelsenbeck et al., 2002; Yang and Rannala, 2005; Alfaro and Holder, 2006). When little is known about the quantity of interest, researchers

[7]Insofar as the most popular method to approximate posterior distributions, Markov chain Monte Carlo, allows for simultaneous estimation of all quantities of interest.

usually resort to "non-informative, ignorance priors". Seaman III et al. (2012), for instance, alert that specifying such "uninformative" priors on model parameters can induce rather informative priors on quantities of interest, specially when these quantities are non-linear functions of the parameters. A characteristic of phylogenetic models is that they tend to be rather complex and some parameters are non-linearly related. See, for instance, Yang (1996) (and Figure 2 therein) for a discussion on the relationship between the Gamma heterogeneity parameter α and the transition/transversion ratio κ under the HKY85 (Hasegawa et al., 1985) model. Rannala et al. (2012) and Wang and Yang (2014) also show that seemingly innocuous priors on the branch lengths can have rather extreme effects on the prior for tree length.

1.3.1 The space of trees

In order to understand the challenges of Bayesian phylogenetics it would be desirable to to have a more rigorous description of the parameter space of interest, that is, the space of time-calibrated phylogenies (TCP). In this section I introduce the necessary technical background for a rigorous characterisation of phylogenetic space. The presentation will follow Semple and Steel (2003) for the general theory and Drummond et al. (2002) and Gavryushkina et al. (2013) for time-trees, with minor adjustments. In this thesis, I am concerned with time-calibrated phylogenies which are binary (bifurcating)[8], rooted, fully-ranked labelled phylogenies.

A rooted binary tree $t \in \mathbb{T}$ on n taxa is a graph $G(\boldsymbol{V}_t, \boldsymbol{E}_t)$ with $2n-2$ edges, $n-1$ internal nodes and n leaf/external nodes, also called taxa – making up a total of $2n-1$ nodes. Each vertex (node) $v \in \boldsymbol{V}_t$ has degree 3, except for a special *root* internal node, denoted ρ, which has degree 2. The set $\boldsymbol{V}_t$ has a partial ordering, defined as follows: $u \preceq v$ if there is a unique simple path from the root ρ to v through u, in

[8]While time-calibrated phylogenies need not be fully resolved (bifurcating), I shall make this simplifying assumption throughout the thesis.

which case we say u is an *ancestor* of v. Denote $\boldsymbol{I}_t = \{x : x \preceq y,\ x, y \in \boldsymbol{V}_t\} \subset \boldsymbol{V}_t$ as the set of interior nodes of t. The root node ρ is then the ancestor of all nodes and the smallest element of the ordering imposed by $\preceq$. The set $\boldsymbol{C}_t = \boldsymbol{V}_t \setminus \boldsymbol{I}_t$ is the set of exterior nodes (taxa) of t.

Let X be a non-empty set of labels, $\phi : X \to \boldsymbol{V}_t$ be a bijective map and $h : \boldsymbol{I}_t \to \{1, 2, \ldots, |\boldsymbol{I}_t|\}$ an (injective) *ranking* function such that $u \preceq v$ implies $h(u) \leq h(v)$ for all $v, u \in \boldsymbol{I}_t$. Notice that $h(u) = h(v)$ implies either $u \equiv v$ or $u, v \in \boldsymbol{C}_t$. A *ranked rooted tree* is an object $t = (\boldsymbol{V}_t, \boldsymbol{E}_t, \rho, \phi, h)$, $t \in \mathbb{F}$. We can supplement t with a set of edge (branch) lengths $\boldsymbol{b} = \{b_1, b_2, \ldots, b_{2n-2}\}$, $\boldsymbol{b} \in \boldsymbol{B} \subseteq \mathbb{R}_+^{2n-2}$, creating a *fully-ranked rooted phylogeny* in the form of the object $(t, \boldsymbol{b}) = \tau \in \boldsymbol{\Psi}$. For convenience, I will henceforth call t a **topology** and τ a **phylogeny**.

It is well known that the cardinality of the space of (partially ranked) rooted topologies on n taxa is $R_n = |\mathbb{T}| = n!(n-1)!/2^{n-1}$. Here, however, we are concerned with *fully ranked* phylogenies, specifically those for which the mapping h is a height function that measures node ages in calendar units. If we associate an age in calendar time with each of the $2n-1$ nodes in a rooted binary tree, these can then be ranked and then used to form a poset $\boldsymbol{a} = \{a_1, a_2, \ldots, a_{2n-1}\} \in \boldsymbol{A} \subset \mathbb{R}_+^{2n-1}$. A convenient labelling is to make labels increase with age, such that $i > j$ implies $a_i \geq a_j$; thus, the root node will have label $2n-1$. An edge $e_{i,j}$ with $i > j$ represents an ancestral lineage and node k in $\boldsymbol{I}_t$ corresponds to a *coalescence* event of two ancestral lineages at time a_k.

It is convenient to define $\boldsymbol{a}_L$ and $\boldsymbol{a}_I$ as the ages of the leaf and internal nodes, respectively. Also denote $\boldsymbol{a} = (\boldsymbol{a}_I, \boldsymbol{a}_L) \in \boldsymbol{A} \subset \mathbb{R}_+^{2n-1}$. There exists a bijective mapping $D : \boldsymbol{B} \to \boldsymbol{A}$ that maps the branch lengths of a TCP to its node ages. In many phylodynamic applications, taxa are sampled through time, leading to *serially-sampled* data sets (Drummond et al., 2002). Hence, here $\boldsymbol{a}_L$ is fixed (for any $\tau \in \boldsymbol{\Psi}$) as it relates to the data collection process. These sampling patterns are important

because they alone impose constraints on the space of phylogenies. Gavryushkina et al. (2013) derive algorithms to count these objects and show that the number of fully ranked phylogenies exceeds R_n. In Figure 1.3 I show the ratio between the number of objects in the space we are interested in ($F_n = |\mathbb{F}|$) and R_n.

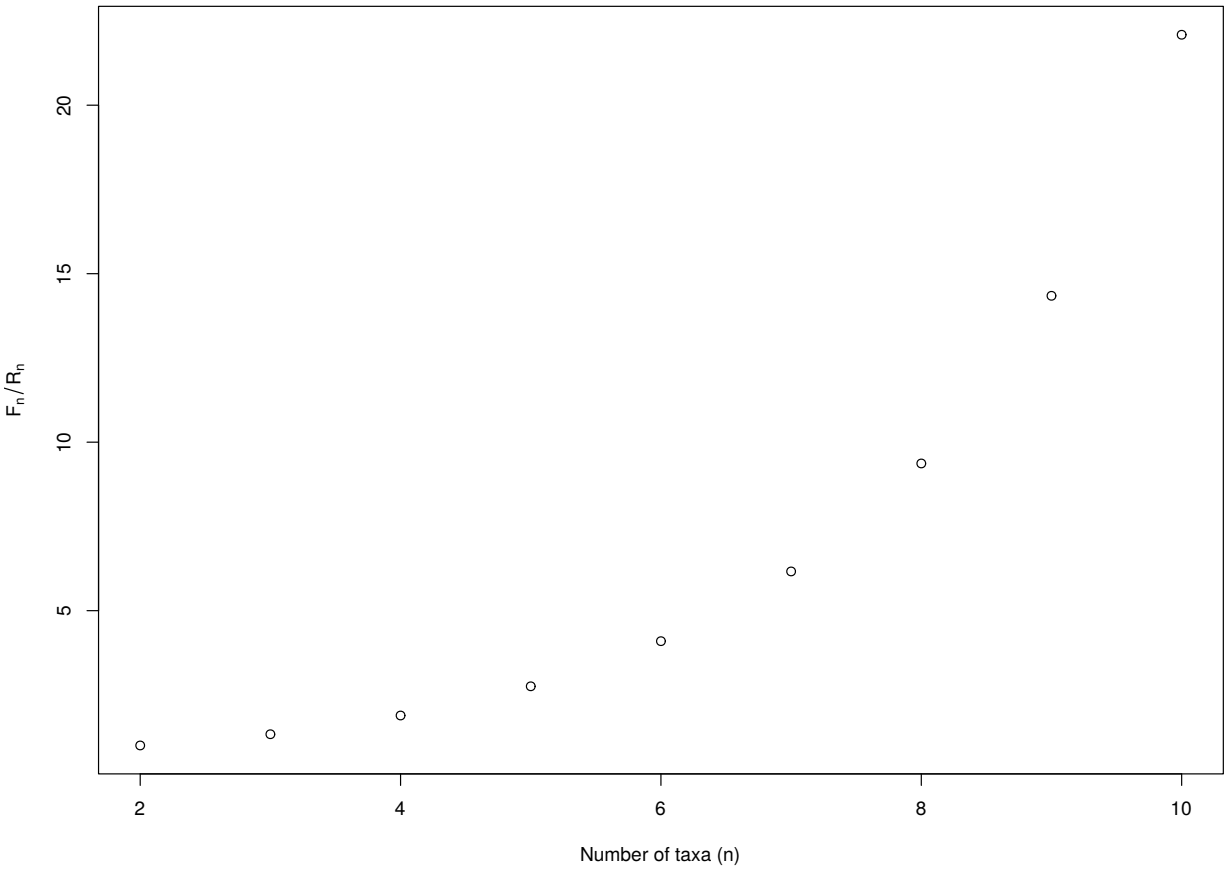


Figure 1.3: Ratio of the number of ranked versus fully ranked trees (with unique sampling times). I show the ratio between the number of of fully-ranked trees on n taxa with n unique sampling times (F_n) and ranked rooted trees (R_n). Numbers extracted from Table 2.2 in Drummond and Bouckaert (2015) which were computed following Gavryushkina et al. (2013).

We are now in position to define the set of **intercoalescent intervals** (also

called divergence times) as $\boldsymbol{s} = \{s_2, s_3, \ldots, s_n\} \in \boldsymbol{S} \subset R_+^{n-1}$, where $s_i = a_{n-i+1} - a_{n-i}$ for $i = 2, \ldots, n-1$ and $s_n = a_1$. Notice that for trees with tips sampled through time, $\boldsymbol{s}$ will have to be slightly adjusted to include subintervals that correspond to the intervals between either a coalescence or sampling event. As explained by Drummond and Bouckaert (2015), the (infinite) space of time-calibrated phylogenies (TCP) can be composed as $\boldsymbol{\Psi} = \mathbb{F} \times \mathbf{S}$, and it is this space I refer to as **phylogenetic space** throughout the thesis.

Finally, let k_i denote the number of existing lineages in the interval $[a_{i-1}, a_i]$, $\boldsymbol{k} = \{k_2, k_3, \ldots, k_n\} \in \boldsymbol{K} \subset \mathbb{N}^{n-1}$. Defining these quantities – intercoalescent intervals and numbers of lineages – is important in that many prior distributions commonly used in Bayesian phylogenetics are based on coalescent processes and the measure of a phylogeny τ depends on it only through its coalescent intervals and numbers of lineages, $\boldsymbol{s}(\tau)$. In Chapter 2 (section 2.2.2) I explore some properties of the posterior under coalescent priors. For serially-sampled phylogenies, the number of coalescent/sampling events (and inter-coalescent intervals), $N_z = |\boldsymbol{s}| = |\boldsymbol{k}|$ is at least $n-1$ and at most[9] $2(n-1)$.

Remark 1.3.1. *Suppose the phylogeny τ has a (fixed) tip sampling structure $\boldsymbol{a}_L$. Consider a phylogeny $\tau^\star$ which is identical to τ except for the fact that it has contemporaneous tips, i.e., $a_i = 0$, $\forall\, a_i \in \boldsymbol{a}_L^\star$, branch lengths being extended accordingly. The mapping $(\boldsymbol{k}^\star, \boldsymbol{a}_L) \to \boldsymbol{k}$ is surjective non-injective.*

Proof. When we have distinct sampling dates, these can be seen as extra nodes that will in turn induce a partitioning of the numbers of lineages. Notice that $\boldsymbol{k}^\star = \{2, 3, \ldots, n\}$ for any bifurcating τ with $N_z^\star = |\boldsymbol{k}^\star| = n-1$, $\sum_{u=2}^{n} k_u^\star = n(n-1)$ and $|\boldsymbol{k}| > N_z^\star$, $\forall\, n \geq 3$. □

[9] This latter case occurs when there are n distinct sampling times for the n tips and these fall exactly in between sampling dates, creating a ladder tree in the same way as the extreme example in the proof of Remark 2.2.4.

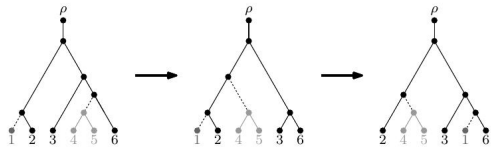
This small observation is useful for us to translate results that hold for SPR random walks on ultrametric trees ($\mathbb{T}$) to serially-sampled ones ($\mathbb{F}$).

SPR graph

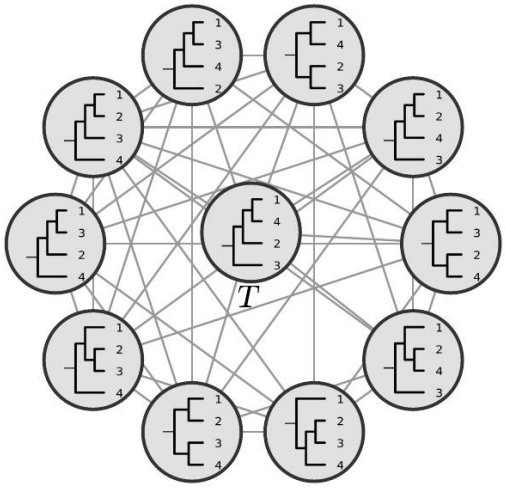
The subtree-prune-and-regraft (SPR) operation picks a node in a tree t, prunes the subtree below that node and regrafts the subtree at a different node in the tree, creating a new tree t' (Figure 1.4A). The SPR distance between two trees x and y, $d_{\mathrm{SPR}}(x, y)$, is then defined as the number of SPR operations needed to obtain y from x or vice-versa. When x and y are rooted trees, we sometimes also call d_{SPR} the rSPR (rooted SPR) distance – this has computational implications, since rSPR is much easier to compute. This distance is also biologically relevant, as it relates to *horizontal gene transfer*, a major evolutionary mechanism in bacteria, for instance.

One way to represent the space of phylogenies is to construct a graph $G_n = (V_n, E_n)$ where each vertex is a tree topology and an edge exists between two vertices if they are "neighbours" in a particular sense. In the SPR graph (Figure 1.4B) two vertices (trees) i and j are connected (neighbours) iff $d_{\mathrm{SPR}}(i, j) = 1$.

This graph-theoretic representation of phylogenetic space is useful in that studying random walks and percolation in the induced graph sheds light into the problem of traversing phylogenetic space. The SPR graph has been studied in some detail. For instance, we know that under an SPR random walk there exist Hamiltonian paths, that is, paths that visit each tree exactly once (Caceres et al., 2011). Whidden and Matsen (2017) employ an optimal transportation approach to define the curvature of the graph in sense of Ricci-Ollivier and show that the SPR graph is flat in the limit. Moreover they show that there might be negative curvature locally, namely for large topological rearrangements – i.e. when two trees differ by one SPR operation that moves a large subtree. In practice this means that there might be "bottlenecks" between trees that differ by large subtree, which can impede efficient traversal of the graph. Finally, Whidden and Matsen (2015) use the SPR graph representation



(a) Two SPR operations



(b) The SPR graph for $n = 4$

Figure 1.4: Subtree prune-and-regraft. Panel A illustrates the subtree prune-and-regraft (SPR) operation: to transform the first tree into the second, the subtree containing $(4, 5)$ is pruned and regrafted at the ancestor of $(1, 2)$. The second SPR operation again prunes $(4, 5)$ and regrafts the subtree at the ancestor of $\{2\}$. In panel B I show the SPR graph for $n = 4$. Figures reproduced from Whidden and Matsen (2017).

to quantify the exploration of phylogenetic space by Markov chain Monte Carlo (MCMC) algorithms. I rely on this representation in the proofs of some assertions throughout the thesis, specially in Chapter 2.

The Billera-Holmes-Vogtmann space

Phylogenetic space does not admit a canonical parametrisation, and hence there are a plethora of representations in the literature. Here I will present the parametrisation

put forth by Billera et al. (2001), since it constitutes a very useful representation for statistical applications (see Willis and Bell (2017) and Dinh et al. (2017)) and also enjoys many desirable theoretical properties (Steel, 2014; St. John, 2017). The first step in constructing the Billera-Holmes-Vogtmann (BHV) space is to associate each possible tree topology with an *orthant* $\boldsymbol{O}_i$, $i = 1, 2, \ldots, F_n$, which corresponds to an embedding of the non-negative real numbers such that $\boldsymbol{O}_i \subset \mathbb{R}_+^{2n-2}$. We are then prepared to construct an *orthant complex* $\boldsymbol{\chi}$ consisting of orthants of dimension $N = 2n - 2$ being joined at the origin and indexed by the countable set Γ such that (a) $\boldsymbol{O}_i \cap \boldsymbol{O}_j$ is a face of both orthants for all $\boldsymbol{O}_i, \boldsymbol{O}_j \in \boldsymbol{\chi}$ and (b) each $x \in \boldsymbol{\chi}$ belongs to a finite number of orthants (Dinh et al., 2017). The orthants in this space are separated by a nearest neighbour interchange (NNI) operation. An schematic representation of the BHV orthant complex space is show in Figure 1.5.

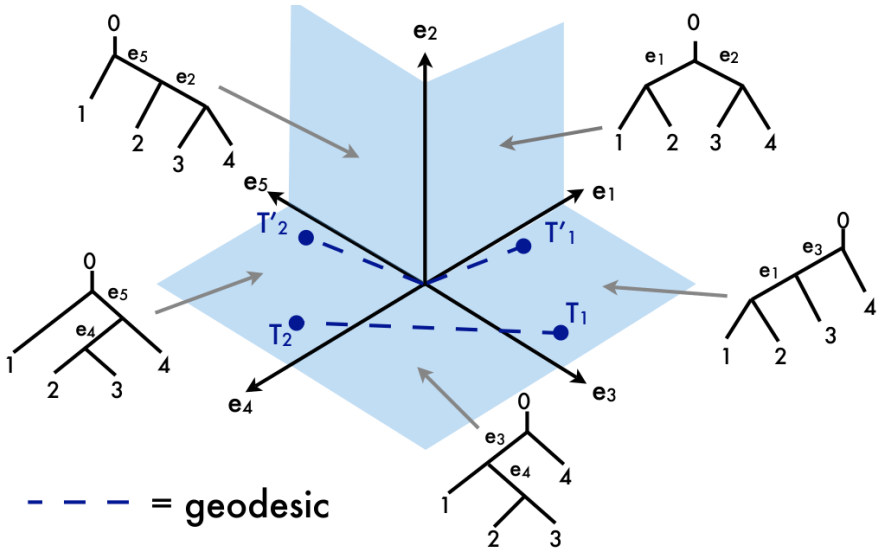


Figure 1.5: Schematic representation of BHV space. Notice how each quadrant is associated with a unique topology t and how some geodesics pass through the origin. Figure extracted from `http://www.crm.umontreal.ca/CanaDAM2009/pdf/owen.pdf` and reproduced with permission.

It can be shown that this construction admits a continuous geodesic metric with non-positive curvature. For two phylogenies with the same tree topology, the BHV

distance can be computed in a straightforward manner by any continuous metric in $\mathbb{R}_+^N$, since both trees lie in the same orthant. When trees have different topologies, one needs to compute the total path length necessary to travel the edge of the orthants. This intuitively accounts for the fact that phylogenies that differ in topology as well as in branch lengths are more different and hence should be more distant. This metric has many desirable properties, including allowing for a unique representation of a "mean" tree, which is useful when the goal is to summarise a set of trees, for instance. Owen and Provan (2011) propose an efficient way of finding the geodesic path between trees, but computing the BHV distance between any two trees remains a computationally hard task, and is substantially more expensive to compute when compared with the Robinson-Foulds (RF) distance (Robinson and Foulds, 1981), for example (see Chapter 3 for definitions of many phylogenetic metrics).

Clade space

For another useful representation of trees, one can also consider bipartitions of the leaves, called "splits" or "clades". An example of a split is $\{A, B, C\}|\{D, E\}$, where A, B, C, D and E are tips. We will refer to the set of all possible splits on n taxa as $\boldsymbol{C}$. For $n \geq 3$ taxa, there are $|\boldsymbol{C}| = 2^{n-1} - 1$ possible splits, and a given tree can contain at most $2n - 2$ splits. Two splits $U_1|V_1$ and $U_2|V_2$ are said to be *compatible* iff at least one of the intersections $U_1 \cap U_2$, $U_1 \cap V_2$, $V_1 \cap U_2$ and $U_1 \cap U_2$ is empty. Let $\Omega(\boldsymbol{C})$ be the σ-algebra of of $\boldsymbol{C}$ and let $\boldsymbol{C}^\star \subset \Omega(\boldsymbol{C})$ be the space of pairwise compatible clades, *i.e.*, S $\boldsymbol{x} \in \boldsymbol{C}^\star$ iff all clades in $\boldsymbol{x}$ are pairwise compatible.

Theorem 1.3.2. *Splits-equivalence (Buneman, 1971): if $\boldsymbol{c}$ is set of pairwise compatible splits, then there exists one and only one topology t that corresponds to $\boldsymbol{c}$. Equivalently: there is a bijective mapping $f : \mathbb{T} \leftrightarrow \boldsymbol{C}^\star$.*

Notice this space is of much smaller cardinality than the space of trees (topologies), which facilitates its analysis in empirical settings (see Section 3.2.4 in chapter 3).

Having described the parameter space of interest, I shall now move on to detail the construction of prior measures on $\mathbf{\Psi}$.

1.3.2 Coalescent models as phylogenetic priors

As discussed above, performing Bayesian inference entails assigning a probability measure over the parameter space. One way of constructing a prior measure on the space of phylogenies is by considering the so-called *coalescent*, described in the pioneering work of Kingman (1982). Coalescent theory seeks to mathematically describe the joining (coalescence) of lineages backwards in time until their common ancestor. The main idea is to relate the effective population size N_e to the probability of any two lineages joining at a certain time t in the past. For instance, in a haploid population with random mating and constant (through time) size of N_e, there are $2N_e$ possible coalescence points (parents) in the previous generation, which leads to there being a probability of $1/2N_e$ that any two lineages coalesce. See below for other models of population dynamics that allow for N_e to change over time. When N_e is sufficiently large, we can approximate the probability that any two lineages coalesce after time t by

$$P(t) = \frac{\exp\left(-(t-1)/2N_e\right)}{2N_e}$$

Parametric tree priors

As seen above, the simplest model one can consider is when N_e has remained constant through time, at size N, $N_e(t) = N$. This population size N hyper-parameter can be given a prior and estimated from the data. This model is suitable whenever the population has remained stable over the time span of the most recent common ancestor of the samples, and provides a baseline to which more parameter-rich models can be compared. One such example of more complex model is the exponential model, which has two parameters: the population size at present N_0

and the growth rate r. The assumption is that population grew exponentially since the time to the most recent common ancestor (tMRCA):

$$N_e(t) = N_0 \exp(-rt),$$

which is suitable to the analysis of early viral samples from epidemics due to initial epidemic growth being approximately exponential.

Let $\tau = (t, \boldsymbol{b})$ be a phylogeny and $\psi(\tau) = (\boldsymbol{s}, \boldsymbol{k})$ be its inter-coalescent intervals and numbers of lineages, respectively. Under a coalescent model with constant (fixed) population size N_e then (Drummond and Bouckaert, 2015, Chapter 2):

$$\pi_0(\tau|N_e) = \prod_{i=1}^{2n-1} \frac{\binom{k_i}{2}}{N_e} \exp\left(-\frac{\binom{k_i}{2} s_i}{N_e}\right), \tag{1.6}$$

$$= \frac{1}{(N_e)^{n-1}} \prod_{j=2}^{2n-1} \exp\left(-k_j(k_j-1)\frac{s_j - s_{j-1}}{2N_e}\right), \tag{1.7}$$

$$\propto \exp\left(-\sum_{j=2}^{2n-1} k_j(k_j-1)(s_j - s_{j-1})\right). \tag{1.8}$$

Non-parametric models

These parametric models might, however, prove too inflexible in approximating population trajectories in real data. Fortunately there are non-parametric models that allow a flexible approach to demographic modelling by constructing a piecewise process which models population size changes between coalescent events (inter-coalescent intervals) (Pybus et al., 2000; Minin et al., 2008; Gill et al., 2012).

The first such model I will consider here is the Skyride model (Minin et al., 2008), which improves on previous semi-parametric models (Pybus et al., 2000) of piecewise population size change by (i) assuming population size changes smoothly over time and (ii) places a smooth Gaussian process prior on the population sizes. Skyride operates on inter-coalescent intervals, i.e., intervals of time between coalescent

events. For a phylogeny with n tips/leaves, let $\boldsymbol{s} = (\boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)$ be the inter-coalescent intervals. If sampling is heterochronous, sampling times further divide inter-coalescent intervals in sub-intervals, i.e., $\boldsymbol{s}_k = (s_{k0}, \ldots, s_{kj_k})$. If we denote the population sizes by $\boldsymbol{\theta} = (\theta_2, \ldots, \theta_n)$, the likelihood becomes

$$Pr(\boldsymbol{s}|\boldsymbol{\theta}) = \prod_{k=2}^{n} Pr(s_k|\theta_k),$$

with

$$Pr(s_k|\theta_k) = \frac{n_{k0}(n_{k0}-1)}{2\theta_k} \exp\left(-\sum_{j=0}^{j_k} \frac{n_{kj}(n_{kj}-1)s_{kj}}{2\theta_k}\right).$$

If we make the convenient transformation $\gamma_k = \log(\theta_k), k = 2, \ldots, n$, we can then place the Gaussian Markov random field (GMRF) prior on $\boldsymbol{\gamma}$:

$$Pr(\boldsymbol{\gamma}|\tau) \propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}\sum_{k=2}^{n-1} \frac{(\gamma_{k+1}-\gamma_k)^2}{\delta_k}\right),$$

where δ_k is the (1d) distance between intervals and τ is the precision parameter associated with the smoothing. For details please see Minin et al. (2008).

The second model I will consider is the Skygrid model, an extension of the Skyride that allows for multiple loci. While in Skyride the estimated trajectory changes at coalescent times, in Skygrid changes occur at pre-specified fixed points in (real) time. This allows population sizes to be estimated for multiple genealogies at once, e.g., when several genes are under analyses and have different genealogies. The researcher must select the number M of grid points to be used and a cut-off K in calendar time. The cut-off K is crucial to the Skygrid analysis, as it is the last point at which population sizes change and hence should be chosen commensurate with the age of the root. As with Skyride, the smoothness of the Skygrid prior is controlled by a precision parameter τ. The Skygrid model presents better statistical properties and is more general, which has led to it superseding Skyride in recent years. These models are parameter-rich and their use is preferable when the data are strongly informative about population history. More recent developments allow for the inclusion of (time)

covariates to inform the population sizes in Skygrid (Gill et al., 2016), which not only promises to help with *understanding* population dynamics but also offers a way of introducing information to inform the non-parametric prior and regularise inference.

Historically, in addition to being used to construct prior distributions these models have been used directly to investigate how distinct patterns of coalescence and genetic diversity relate to population epidemiological processes, specially for viruses (Rodrigo and Felsenstein, 1999; Pybus et al., 2000; Pybus and Rambaut, 2009). Examples include the analysis of the dynamics of Hepatitis C virus (HCV) in Egypt by Pybus et al. (2003) which found that "...the Egyptian HCV epidemic was initiated and propagated by extensive antischistosomiasis injection campaigns" and Rambaut et al. (2008) who analysed over 1300 Influenza virus complete genomes and found differences in the evolutionary dynamics of the A/H3N2 and A/H1N1 subtypes by spotting differences in their skyline plots (Figure 1 therein).

Other priors

While the coalescent offers a flexible framework for modelling the population dynamics from genealogies, it also relies on restrictive assumptions. For instance, the coalescent assumes that the fraction of sampled individuals (number of taxa, n) is a negligible fraction of total population (Fu, 2006; Volz et al., 2009). With the rapid increase in the number of sequences, it is quite possible that most cases in an epidemic could be sampled, therefore rendering this assumption problematic. To address this and also allow for more explicit models of epidemic dynamics, several models based on birth-death processes, Volz et al. (2009), Rasmussen et al. (2011) and Stadler et al. (2011) have developed approaches that incorporate other population models such the Yule process and epidemic models such as the Susceptible-Infected-Removed (SIR) model. Similarly to the coalescent, these

models too can be used as prior measures for phylogenies, and in addition be used to estimate parameters of interest such as the basic reproductive number, R_0.

1.4 Markov chain Monte Carlo

Since the phylogenetic posterior (1.5) is not available in closed-form even for the simplest models, it must be numerically approximated. In the following sections I introduce the necessary mathematical background for Markov chain Monte Carlo (MCMC), which I will base mostly on Geyer (2011). For ease of exposition, whenever a choice is to be made between a discrete and a continuous setting I shall assume the latter – as discussed above, phylogenetic space has both discrete and continuous components.

Suppose one aims to sample from a distribution[10] $\pi(\cdot)$ defined on a sample space $\mathcal{X}$, with density π_d such that $\forall U \subseteq \mathcal{X}$

$$\pi(U) = \frac{\int_U \pi_d(x)dx}{\int_{\mathcal{X}} \pi_d(x)dx}.$$

One reason to obtain samples from $\pi(\cdot)$ is to compute expectations of (Borel-measurable) functionals $g : \mathcal{X} \to \mathbb{R}$ such that[11]

$$\mu_g := \mathbb{E}_\pi[g(X)] = \int_{\mathcal{X}} g(x)\pi_d(x)dx.$$

Classic Monte Carlo theory says that, under some mild regularity conditions, if one obtains a sample of i.i.d. random variables $\boldsymbol{Z} = \{Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}\} \sim \pi$, then

$$\widehat{\mu_g} = N^{-1} \sum_{i=1}^{N} g(Z^{(i)}),$$

[10] I will assume the target is either normalised or can be normalised, *i.e.*, $0 < \int_{\mathcal{X}} \pi_d(x) < \infty$.
[11] It is important to note that π_d is normalised in this setting.

is an unbiased estimate of μ_g. Moreover, the standard deviation of $\widehat{\mu_g}$ is $\mathcal{O}(\frac{1}{\sqrt{N}})$, meaning we can make the estimate as precise as desired by increasing the number of samples N. Obtaining $\boldsymbol{Z}$ directly, however, might be impractical. Instead, Markov chain Monte Carlo (MCMC) is a technique to draw samples by constructing a Markov chain $\{X_i\}$ on $\mathcal{X}$ that has $\pi(\cdot)$ as its *limiting* (or stationary) distribution. More formally, we want to construct $\{X_i\}$ with transition probabilities $P(x, dy)$ such that

$$\pi(dy) = \int_{\mathcal{X}} \pi(dx)P(x, dy)$$

for all $x, y \in \mathcal{X}$. If we are able to draw samples from this Markov chain, then for a sufficiently large number of samples we will obtain a collection of random variables that are approximately drawn from $\pi(\cdot)$.

We are still however left with the task of finding an appropriate $P(x, dy)$. One useful simplifying assumption usually made is that $P(x, dy)$ is *reversible*:

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx).$$

This condition is also known as **detailed balance** and ensures $\{X_i\}$ has $\pi(\cdot)$ as its stationary (limiting) distribution. One of the simplest ways of constructing a reversible Markov is the so-called Metropolis-Hastings algorithm, described in Section 1.4.1.

1.4.1 Metropolis-Hastings

Let $q_\sigma(x, y)$ be a *candidate-generating* density with indexing parameter σ such that $\int_{\mathcal{X}} q_\sigma(x, v)dv = 1 \forall x \in \mathcal{X}$. Now consider a Markov chain $Q_\sigma(x, \cdot)$ such that $Q_\sigma(x, y) \propto q_\sigma(x, y)dy$. The so-called Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) consists of constructing a Markov chain with acceptance probability

$$\alpha_\sigma(x, y) = \min\left[1, \frac{\pi_d(y)q_\sigma(y, x)}{\pi_d(x)q_\sigma(x, y)}\right], \tag{1.9}$$

with $\alpha_\sigma(x, y) = 1$ if $\pi_d(x)q_\sigma(x, y) = 0$ (Chib and Greenberg, 1995, p. 329). The quantity $q_\sigma(y, x)/q_\sigma(x, y)$ is called the "Hastings ratio" and acts as a correction factor to ensure we sample from the desired target (see below).

Computationally, MH can be described as generating samples $\boldsymbol{Z}$ as follows:

0. Pick some $Z^{(0)} \in \mathcal{X}$ with $\pi_d(Z^{(0)}) > 0$;

   for $n = 0$ to M:

1. Given $Z^{(n)}$, generate a *proposal* $Y^{(n+1)} \sim Q_\sigma(Z^{(n)}, \cdot)$;

2. Sample $u \sim \text{Uniform}(0, 1)$;

3. If $\alpha_\sigma(Z^{(n)}, Y^{(n+1)}) > u$, set $Z^{(n+1)} = Y^{(n+1)}$, otherwise $Z^{(n+1)} = Z^{(n)}$;

This can be shown to correctly sample from $\pi(\cdot)$ for appropriately chosen $q_\sigma(\cdot, \cdot)$ (see below) and is straightforward to implement on a computer. The MH algorithm is a very popular workhorse of MCMC, much due to its simplicity of implementation. Good introductions can be found in Chib and Greenberg (1995) and Robert (2015).

While the MH algorithm is by far the most popular algorithm in MCMC, it is by no means the only one. A very popular method is the Gibbs sampler (Geman and Geman, 1984), an algorithm whereby samples from $\pi(\cdot)$ are drawn from a series of conditional distributions, which can be quite useful when sampling Bayesian posterior distributions based on conjugate priors. When sampling from spaces with varying dimension, the so-called "Reversible-jump" MCMC of Green (1995) has enjoyed great success. Perhaps not surprisingly, many of these algorithms are special instances of a more general sampling procedure, as shown by Keith et al. (2004). A common criticism of MH is the fact that the transitions (step 3 above) do not take into account the structure of the target, often leading to random-walk behaviour. Algorithms such as the Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Rosenthal, 1998) and Hamiltonian (Hybrid) Monte Carlo (HMC) (Duane et al., 1987; Neal et al., 2011) exploit the structure of the target – usually in the form of

gradients – to perform guided transitions and increase efficiency. I discuss this point further in Chapter 6.

1.4.2 Transition kernels

Successful exploration of the target distribution depends crucially on the choice of $Q_\sigma(\cdot,\cdot)$. In particular, the **transition kernel**

$$\kappa_\sigma(x, dy) = q_\sigma(x, y)\alpha_\sigma(x, y)dy + \left[1 - \int_{\mathcal{X}} q_\sigma(x, y)\alpha_\sigma(x, y)dy\right]\delta_x(dy), \qquad (1.10)$$

where $\delta_x(u) = 1$ if $x \in u$ and 0 otherwise, needs to be constructed carefully to ensure correct and efficient sampling. In section 1.4.3 I briefly discuss some of the aspects, both practical and theoretical, involved in designing effective transition kernels. Further discussion of transition kernels in MCMC for Bayesian phylogenetics can be found in Chapters 2 and 6.

For complex parameter spaces, the construction of the proposal might be quite complicated, and the the density $q_\sigma(\cdot,\cdot)$ might be hard to compute. Green (2003) proposes a constructive method for computing the acceptance ratio that simplifies calculations, specially when considering varying-dimension problems. Denote the current state of the chain by x and the proposed (candidate) state by x'. The main idea is to sample a set of auxiliary random variables $\boldsymbol{u} \in \boldsymbol{U}$ from a distribution $g(\cdot)$ and then apply a transformation $h : \mathcal{X} \times \boldsymbol{U} \to \mathcal{X}$ such that $x' = h(x, \boldsymbol{u})$. To obtain x from x' the procedure is to generate a set $\boldsymbol{u}' \in \boldsymbol{U}$ from a (potentially different) distribution $g'(\cdot)$ and apply a transform $h' : \mathcal{X} \times \boldsymbol{U} \to \mathcal{X}$ to get $x = h'(x', \boldsymbol{u}')$. The acceptance probability can then be written as

$$\alpha^\star_\sigma(x, y) = min\left[1, \frac{\pi_d(y)g'_d(\boldsymbol{u}')}{\pi_d(x)g_d(\boldsymbol{u})}|J|\right],$$

with $J = \det\left[\frac{\partial(x',\boldsymbol{u}')}{\partial(x,\boldsymbol{u})}\right]$. See Holder et al. (2005) for an application of this technique

in Bayesian phylogenetics, where it was used to correct an error in the original derivation of the `LOCAL` transition kernel (Larget and Simon, 1999).

1.4.3 General considerations on MCMC

In practice, there are many details that determine the effectiveness of MCMC as a numerical method. One of the key challenges in efficiently computing an approximation of the target is the boldness of the proposing mechanism: propose extreme (bold) values and the candidate is likely to be rejected, making the chain be stuck for long periods; propose conservatively and the candidate is likely to be accepted but the chain will move very little away from the current state, leading to poor exploration. It is therefore desirable to construct the candidate generating density so as to strike the best balance between bold and conservative proposals, a problem called optimal scaling (Roberts and Rosenthal, 1998; Roberts et al., 2001). Gelman et al. (1996) suggest that under some regularity conditions on the target $\pi(\cdot)$, an acceptance probability of 0.234 when the dimension of $\mathcal{X}$ is large enough will produce optimal sampling when compared with independent samples from $\pi(\cdot)$. More modern algorithms are **adaptive**, meaning the transition kernel indexing parameter σ can be "tuned" using the previous history of the chain to achieve the optimal acceptance probability (Haario et al., 2001).

Another issue that has received much attention in the literature is determining whether $\{X_i\}$ has *converged in distribution* to $\pi(\cdot)$. See Robert and Casella (2004) and Meyn and Tweedie (2012) for the technical details and Chapter 3 for a review on (statistical) tools to assess convergence. Related to the issue of convergence is ascertaining that the effect of the initial state $Z^{(0)}$ is negligible. In actual practice, approximations to functionals are usually done considering only samples after a **warm-up** (burn-in) number of iterations W, *i.e.*, $\hat{\pi}(f) = (M-W)^{-1}\sum_{i=W+1}^{W+M} f(Z^{(i)})$ (see Chapter 3). We are usually interested on **mixing**, that is, how well the Markov chain explores the target distribution. Since the samples

obtained by MCMC are fundamentally correlated, we can measure its *efficiency* by estimating how many independent samples from $\pi(\cdot)$ were produced. This quantity is called the **effective sample size** (ESS) and I provide definitions and discussion in Chapter 3. In short, the closer the ESS is to the chain length[12] M, the more efficient is MCMC in approximating $\pi(\cdot)$. In practical applications one also has to consider computational ("wall clock") performance, *i.e.*, how fast a particular method produces a given sample, measured for instance in ESS/hour.

In addition to efficiency it is important to ensure comprehensive exploration of the target distribution. In order for samples obtained with MCMC to be of any practical use, they need to represent the **typical set** of the target distribution. Intuitively, this is to be understood as exploring the "bulk" of the distribution, where most of its mass lies. More formally, the typical set can be defined as $\boldsymbol{S}_\epsilon \in \mathcal{X}$ such that for all sequences $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\} \in \boldsymbol{S}_\epsilon$

$$2^{-n(H_\pi+\epsilon)} \leq \pi(\boldsymbol{X}) \leq 2^{-n(H_\pi-\epsilon)}$$

holds, where $H_\pi = \mathbb{E}_\pi[\log_2(x)]$ is the entropy of $\pi(\cdot)$. In practice we cannot know for sure whether our chain has explored the typical set. Even if our algorithm is correctly constructed and our efficiency measures suggest satisfactory sampling, it might be the case that our chain is stuck at a mode, for instance, and not sampling the whole space. For most target distributions of interest in practice, the issue of determining the typical set remains an open problem. See chapter 3 for a first – and informal – stab at characterising the typical set for time-calibrated phylogenies.

Contrary to what is routinely suggested in the literature, MCMC is not a Bayesian method. Rather, MCMC is a computational method for approximating integrals and is completely agnostic about what is being computed. As a counter-example to the claim the MCMC is "Bayesian" in any way, I offer Geyer (1991) and Kuhner

[12] It is technically possible to have ESS larger than M when using over-relaxation techniques, but this is a fringe case.

et al. (1998), who employ MCMC for maximum likelihood estimation[13]. Moreover, MCMC is not the only tool to compute integrals and approximate distributions; particle filtering – also called sequential Monte Carlo –(Gordon et al., 1993; Del Moral, 1996) –, rejection sampling (Casella et al., 2004) and importance sampling (Rubinstein and Kroese, 2016) are examples of techniques that do not involve the construction of Markov chains.

1.5 Software

In the course of my doctoral work I have written a fair amount of computer programs (scripts) to perform several custom analyses as well as to automate data wrangling and processing tasks. In this short section I detail some of the software I have used in my research and acknowledge the efforts of the many programmers who developed these tools without which my work would not have been possible. The first point to notice is that most (if not all) software I have used was open source software, that is, software for which the source code is released under a licence that allows the end user to study, change, and distribute the software to anyone and for any purpose (Laurent, 2004). This is important insofar as it allows one to see what is "under the hood", the inner workings of the software under use, a crucial aspect of ensuring scientific correctness (Darriba et al., 2018). Throughout my PhD I have used a GNU/Linux desktop (and server) system, which sports a host of useful tools that were crucial for my research. The first of these tools is the bash shell (from UNIX), which allowed me to script many otherwise tedious and error-prone tasks. A second tool that I made extensive use of was `GNU Parallel` (Tange, 2011), which allows executing jobs in parallel and hence exploit multi-core architectures in most modern personal computers and servers.

[13] See also `http://users.stat.umn.edu/~geyer/mcmc/diag.html` – also cited in Chapter 3 – for the precious quote: "[MCMC] isn't even statistics, it is a tool.". Terms such as "Bayesian MCMC" make no sense whatsoever.

Most of the programs I wrote for my research were written in the R statistical computing language (R Core Team, 2017), using the open source integrated development environment (IDE) Rstudio (RStudio Team, 2015). R offers an enormous variety of user-contributed packages which greatly expand the capabilities of the base distribution. I have used the **ape** (Paradis et al., 2004), **parallel** (R Core Team, 2017), **ggplot2** (Wickham, 2009), **viridis** (Garnier, 2018) and **data.table** (Dowle and Srinivasan, 2017) packages repeatedly across several projects. This list is by no means exhaustive; please see the individual chapters for more details on the R packages (and other software) I have used in each of the projects. BEAST (see below) is written in the cross-platform JAVA language (`https://java.com/`) and I have used the Eclipse IDE (`https://www.eclipse.org/ide/`) to aid both code reading and minor development. Finally, I would like to mention that I have made extensive usage of the excellent version-control system `git` and the associated code hosting service GitHub (`https://github.com/`). This thesis was written using the open-source typesetting language LaTeX and its source code is hosted at `https://github.com/maxbiostat/PhD_Thesis`.

BEAST

The software package BEAST, short for **B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees (Drummond and Rambaut, 2007; Drummond et al., 2012; Suchard et al., 2018) is a central part of this thesis. BEAST is specially designed for the estimation of rooted phylogenies from genetic sequences, with a heavy focus on time-calibrated phylogenies. It offers a large number of tree priors (coalescent- and non coalescent-based models) as well as specialised computational machinery for the construction and estimation of discrete and continuous phylogeographic models (Lemey et al., 2009, 2010; Pybus et al., 2012; Dudas et al., 2017). The interested reader is referred to the on-line documentation (`http://beast.community/`) for more information.

With regard to MCMC, BEAST uses Metropolis-Hastings as its main algorithm, while specialised samplers do exist for some model components. For example a Gibbs sampler is used to efficiently sample the quantities in the Skygrid model described above. In BEAST *parlance*, transition kernels are called **operators** and I shall employ the terms interchangeably. BEAST employs adaptive MCMC, meaning that it tunes the scale of its operators to achieve a target acceptance probability – currently 0.234 for most operators. It is important to notice, however, that not all operators in BEAST can be adjusted in this way, and are therefore called "non-tunable", in contrast to the "tunable" ones.

The transition kernels described in Chapter 2 were implemented in BEAST[14] and the visualisations in Chapter 3 would not have been possible without specialised classes in the BEAST code base (see Section 3.3 therein). Moreover, BEAST output was also important in the analyses I present in Chapters 4 and 5, where I sought to combine phylodynamic data from BEAST analyses with epidemiological information.

1.6 Goals

Since its emergence in the later half of the 2000s, phylodynamics has grown into a powerful tool in the study of pathogen dynamics. However, as many authors (Volz et al., 2013; Pybus et al., 2013) note, there is an ever growing gap between data accumulation and the methodological apparatus to analyse and integrate this data.

This PhD thesis is an attempt at plugging that gap, and has two main axes: (a) the development of more efficient transition kernels and visualisations for MCMC in phylogenetic space and (b) the application of state-of-the-art statistical methods

[14] I would like to explicitly acknowledge Andrew Rambaut's extensive help with understanding BEAST and its inner workings, as well as implementing many of the ideas explored in this thesis, for which I am very grateful.

to address phylodynamic/epidemiological questions. In chapters 2 and 3 I outline a proposal for new transition kernels (operators) and address the modification of existing convergence diagnostics to large, time-calibrated phylogenies. Chapters 4 and 5 showcase how one can combine modern statistical tools and phylogenetic/phylodynamic data to explore the evolutionary and epidemiological dynamics of pathogens such as Ebola virus. In chapter 6 I discuss the overall impact of the findings in this thesis and how I see the field moving forward.

Chapter 2

Adaptive transition kernels for Bayesian phylogenetics

> Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse.
>
> Alan Sokal (1955-) in *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms* (1996).

2.1 Introduction

In Bayesian phylogenetics one is usually interested in computing the posterior distribution

$$p(t, \boldsymbol{b}, \boldsymbol{\theta}|D) = \frac{f(D|t, \boldsymbol{b}, \boldsymbol{\theta})\pi(t, \boldsymbol{b}, \boldsymbol{\theta})}{\sum_{t_i \in \mathbb{F}} \int_{\boldsymbol{B}} \int_{\boldsymbol{\Theta}} f(D|t_i, \boldsymbol{b}_i, \boldsymbol{\theta})\pi(t_i, \boldsymbol{b}_i, \boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{b}_i} \,, \tag{2.1}$$

where D is observed data and $t \in \mathbb{F}$ is a fully-ranked tree topology associated set of branch lengths $\boldsymbol{b}$. Finally $\boldsymbol{\theta}$ is a set of parameters such as substitution model parameters, migration rates, heritability coefficients, etc. In many applications, the aim is to construct time-calibrated phylogenies, i.e. phylogenetic trees whose branch lengths are measured in units of calendar time. In particular, one might have sequences sampled through time (heterochronous/serially-sampled) which enable direct estimation of the rate of evolution and reconstruction of past population dynamics (Drummond et al., 2002, 2005). These types of data sets pose additional challenges to inference because they impose constraints[1] on the space of valid trees (Stadler and Yang, 2013).

2.1.1 MCMC in phylogenetic space

One of the main features of the Bayesian approach is to allow parameter inference and hypothesis testing whilst accommodating phylogenetic uncertainty (Suchard et al., 2001; Huelsenbeck et al., 2002; Lemey et al., 2014; Cybis et al., 2015; Baele et al., 2015, 2017). This treatment of uncertainty is achieved by integrating (marginalising) over the space of phylogenies, a task which depends crucially on efficiently traversing tree space. Even for the simplest models, the distribution in (2.1) cannot be computed analytically requiring numerical approximation, usually accomplished through Markov chain Monte Carlo (MCMC). The use of MCMC for Bayesian methods in phylogenetics has grown steadily since its introduction in the late 1990s and early 2000s (Kuhner et al., 1995; Sinsheimer et al., 1996; Rannala and Yang, 1996; Yang and Rannala, 1997; Mau et al., 1999; Li et al., 2000; Suchard et al., 2001; Drummond et al., 2002), with software packages such as Mr Bayes (Ronquist et al., 2012) and BEAST (Drummond et al., 2012; Suchard et al., 2018) becoming widely used by researchers in a broad range of disciplines (Murphy et al., 2001; Bouckaert et al., 2012; Lemey et al., 2014).

[1] More specifically temporal precedence constraints.

The Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) is a very popular MCMC technique due to its generality and ease of implementation. For ease of presentation, let $\tau = \{t, \boldsymbol{b}\}$ be a phylogeny and $q_\gamma(\tau'|\tau)$ be a conditional distribution indexed by a parameter γ from which a new state τ' can be proposed from a current state τ. We call $q_\gamma(\tau'|\tau)$ a **candidate-generating distribution**. It can be shown that accepting/rejecting a new state τ' based on the ratio[2] $A_\gamma(\tau|\tau') = min\left(1, \frac{p(\tau'|D)q_\gamma(\tau|\tau')}{p(\tau|D)q_\gamma(\tau'|\tau)}\right)$ leads to the desired distribution for a suitably constructed proposal mechanism.

There are no "default" choices for the candidate-generating distribution $q_\gamma(\cdot|\cdot)$; it must be chosen with the target distribution (in this case the Bayesian posterior) in mind. Moreover, the efficiency of MCMC algorithms in approximating the target distribution depends crucially on the choice of transition kernel (Brooks et al., 2003; Al-Awadhi et al., 2004; Yang and Rodríguez, 2013; Thawornwattana et al., 2017). As argued by Höhna and Drummond (2012), tree transition kernels are usually built in a relatively simplistic fashion, which in turn leads to inefficient exploration of tree space. Moreover, most transition kernels proposed to date are not adaptive, i.e., the parameter(s) γ cannot be adjusted during the Markov chain to achieve a desired acceptance probability (Haario et al., 2001). Given the clear advantage of adaptive MCMC over non-adaptive implementations for high-dimensional target distributions (Roberts and Rosenthal, 2009; Baele et al., 2017), the development of adaptive tree transition kernels could lead to substantial gains in performance.

Höhna et al. (2008) developed new "clock-constrained" transition kernels to improve efficiency when dealing with time-calibrated trees. The authors point out that clock-constrained trees impose additional restrictions on the state space of the MCMC algorithm and hence that performance could be increased by developing transition kernels that took the extra information provided by tip dates. They develop two such kernels: Fixed node-height Prune-and-Regraft (FNPR) and Intermediate Exchange

[2] I will drop the dependence of the posterior on $\boldsymbol{\theta}$ for convenience of notation. One should however keep in mind some parameters may be strongly correlated with the phylogeny τ.

(IE). FNPR finds a "target" node (excluding the root and its two children) at random, prunes it and regrafts the resulting subtree at a "destination" node in the tree at the same height at random. Intermediate Exchange is similar in spirit, but the regraft node is not chosen uniformly. Instead, IE is constructed to prefer local rearrangements, by picking closer nodes with a higher probability (see below and section II in Höhna et al. (2008) for details). A limitation of FNPR nor IE is that they are not adaptive.

Höhna and Drummond (2012) explored more sophisticated "guided" tree transition kernels, inspired by Gibbs sampling (Geman and Geman, 1984). The idea behind their "Metropolised" Gibbs samplers is to maximise transition probability, i.e., the probability that the chain moves to a new state. This is accomplished by prohibiting the current state as a proposed state, leading to a transition probability of one. The kernels developed in Höhna and Drummond (2012) use a weighting scheme based on conditional clade probabilities (CCP) to guide transitions between trees. A move to a tree with a lower CCP score is thus less likely, whilst a move that increases the score has a higher probability of being accepted. A limitation of these Metropolised transition kernels is that CCP scores require normalisation over all trees (Larget, 2013) and hence can be cumbersome to calculate.

More recently, Dinh et al. (2016) and Fourment et al. (2017) develop "guided" candidate-generating mechanisms for on-line Bayesian phylogenetics via sequential Monte Carlo, which makes use of a surrogate function to make computations feasible. While that is an active field of research, it relies on machinery not yet extended to the time-calibrated case. Hence in this thesis and chapter I will focus on extending existing MH algorithms, which I will argue strikes a balance between computational efficiency and tractability.

To achieve maximum efficiency, a tree transition kernel needs to have the following characteristics: (i) be computationally cheap; (ii) be adaptive; (iii) traverse phylogenetic space quickly, i.e., have a high *mixing rate*. In this chapter I develop and study

two simple, adaptive time-tree transition kernels, which we implement in the open source software package BEAST (`https://github.com/beast-dev/beast-mcmc/`). I analyse performance in real as well as simulated data sets, focusing on the inference of time-calibrated phylogenies. Technical details necessary to make some of the claims in this chapter precise are given in Section 2.2.3.

2.2 New time-tree transition kernels

In this section I introduce two new candidate-generating densities that attain the properties discussed above while being specially designed for time-calibrated phylogenies.

2.2.1 Preliminaries

I briefly lay out some notation and background that will be necessary for the presentation, with further details and more rigorous definitions already given in Chapter 1 (Section 1.3.1). Throughout this chapter I will use $\mathbf{\Psi}$ to denote the parameter space encompassing topologies and branch lengths, henceforth called "phylogenetic space", and $\tau \in \mathbf{\Psi}$ to denote a bifurcating, rooted tree with branch lengths on n taxa[3]. Let $\mathbf{B}(\tau) = \{b_1, b_2, \ldots, b_{2n-2}\}$ be the set of branch lengths of τ. It is possible to compute the height of each node using a mapping $h(\cdot)$ such that $\mathbf{H}(\tau) = \{h_1, h_2, \ldots, h_{2n-1}\}$ is the set of node heights for all nodes (internal and external) in τ. Here we will define $h(i) = l_{max} - l_i$ and $l_i = \sum_{k \in \mathbf{W_i}} b_k$, where $\mathbf{W_i}$ is the minimal path between node i and the root ρ and l_{max} is the maximum of these shortest paths. It is also convenient to define P_i, G_i and S_i as the parent, grandparent and sibling of node i, respectively. With this notation in mind, call $p_i = h(P_i) - h(i)$, $g_i = h(G_i) - h(P_i)$ and $q_i = h(P_i) - h(S_i)$ the corresponding branch lengths. The most recent common ancestor of nodes i and j, mrca_{ij} is the first node

[3]What Drummond and Bouckaert (2015) call a "fully ranked" tree.

in which the paths $\mathbf{W_i}$ and $\mathbf{W_j}$ intersect. Finally, let $\Delta_{ij} = 2h(\text{mrca}_{ij}) - [h(i) + h(j)]$ be the patristic (path) distance between nodes i and j on the phylogeny.

SubTreeJump

The first transition kernel I propose, *SubTreeJump* (STJ)[4], is similar in spirit to both Intermediate Exchange and FNPR. STJ extends IE by introducing a tuning parameter α that controls how local rearrangements are. The idea is to make the probability of moving node i to j proportional to Δ_{ij}^{α} instead of Δ_{ij} as in Intermediate Exchange. The necessary steps to perform a SubTreeJump operation on τ in order to propose a phylogeny τ' are described in detail in Box 1 and an illustration is provided in Figure 2.1. This allows one to favour bold moves away from i ($\alpha > 0$) or more conservative moves closer to i ($\alpha < 0$). Notice that SubTreeJump is equivalent to FNPR for $\alpha = 0$. One can then define the SubTreeJump candidate-generating density as $q_\alpha(\tau'|\tau) = Pr(i \to j)$.

Algorithm 1: SubTreeJump transition kernel.

0 Excluding the root and its children, pick a node i in τ uniformly at random, i.e., with probability $1/(2n-4)$;

1 Determine P_i and compute $h(P_i)$;

2 Construct the set of destination nodes $\mathbf{D_i} = \{d : h(d) \leq h(P_i) < h(P_d)\}$;

3 For all $k \in \mathbf{D_i}$, compute l_{ik};

4 For some fixed $\alpha \in \mathbb{R}$, pick a node j with probability $Pr(i \to j) = \frac{\Delta_{ij}^{\alpha}}{\sum_{d \in \mathbf{D_i}} \Delta_{id}^{\alpha}}$;

5 Prune the phylogeny at P_i and regraft the resulting subtree at P_j, creating a new phylogeny τ'.

Note that while for $\alpha = 0$ the move is symmetric (Höhna et al., 2008), for $\alpha \neq 0$ the density $q_\alpha(\cdot|\cdot)$ is not, since for two arbitrary nodes i and j whilst the sets $\mathbf{D_i}$ and $\mathbf{D_j}$ coincide ($|\mathbf{D_i}| = |\mathbf{D_j}|$), the distances between nodes usually do not. Hence one

[4] An alternative name for STJ is generalised fixed node height prune and regraft, gFNPR.

needs to compute the *Hastings ratio* $q_\alpha(\tau|\tau')/q_\alpha(\tau'|\tau)$. To get a reverse move from τ' back to τ, one needs to pick the original target node i – which is guaranteed to exist in $\mathbf{D_j}$ – as a destination. The Hastings ratio is then[5]

$$\begin{aligned}\frac{q_\alpha(\tau|\tau')}{q_\alpha(\tau'|\tau)} &= \frac{\Delta_{ij}^\alpha}{\sum_{d'\in\mathbf{D_j}}\Delta_{id'}^\alpha}\Big/\frac{\Delta_{ij}^\alpha}{\sum_{d\in\mathbf{D_i}}\Delta_{id}^\alpha}, \\ &= \frac{\sum_{d\in\mathbf{D_i}}\Delta_{id}^\alpha}{\sum_{d'\in\mathbf{D_j}}\Delta_{jd'}^\alpha}. \end{aligned} \tag{2.2}$$

I note two limitations of this transition kernel. First, because $\mathbf{H}(\tau)$ is ultimately discrete for any given τ, not any acceptance probability is attainable. This "granularity" of the kernel density causes problems for most adaptation schemes because the dependence of the acceptance probability on the indexing parameter is implicitly assumed to be smooth (see Section 2.6.1). Secondly, SubTreeJump on its own does not necessarily induce an irreducible Markov chain on the space of rooted topologies – see Section 2.2.3 for a counterexample/proof. This can be easily remedied by combining STJ with branch length candidate-generating mechanisms. Since STJ (and FNPR) does not change node heights or numbers of lineages, it does not change the density under the coalescent prior, *i.e.* $\pi(\tau) = \pi(\tau')$. Whether this is a desirable property is an open question. The disconnect between proposals in topological space and branch length space could be undesirable due to it failing to account for the dependence between topology and branch lengths.

SubTreeLeap

Ideally, we would like to have a single adaptive transition kernel to update topology and branch lengths simultaneously. To this end I propose *SubTreeLeap* (`STL`), a transition kernel based on patristic distances. The central idea behind `STL` is to

[5]Note that I omit the node-picking probability $1/(2n-4)$ because it is the same for both moves. If one were to devise a kernel with non-uniform node-picking probabilities, these would have to be included in the Hastings ratio.

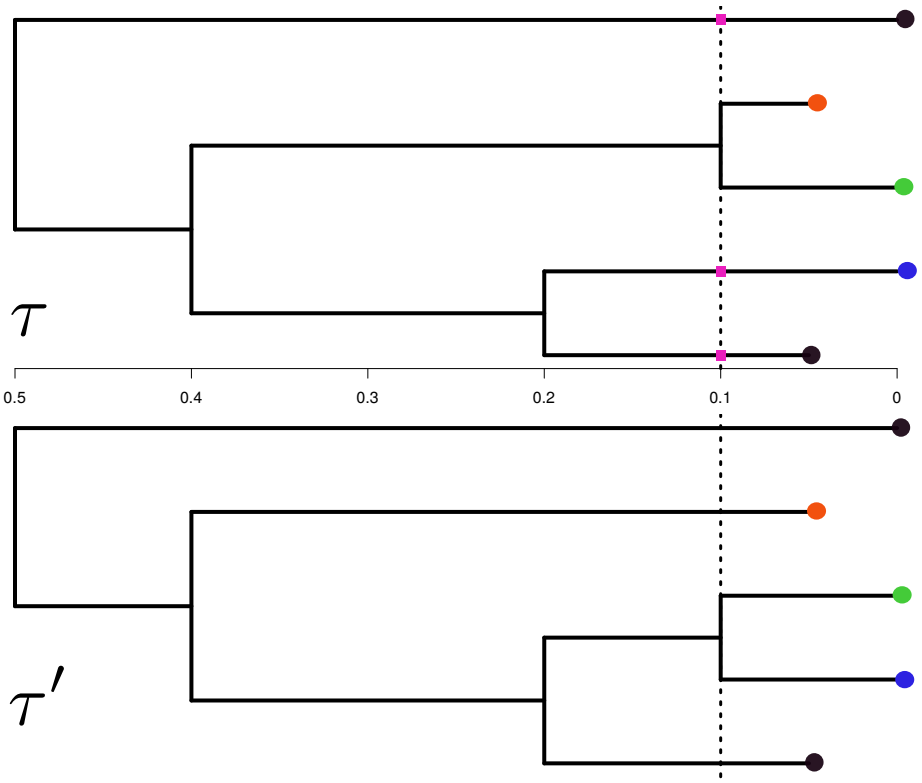


Figure 2.1: Schematic representation of a SubTreeJump proposal. The target node (green) is picked, we find all the nodes (or positions along branches) that intersect at the height of its parent (pink squares) and then a destination is picked with probability as described in the text. Other nodes are coloured so the reader can easily identify the changes.

move a node i to new location in the tree that is at most at (patristic) distance δ from i. To this end, one first draws the distance δ from a distribution $\kappa(\delta|\sigma)$ indexed by a parameter σ, henceforth called the *distance kernel*. One then finds the set $\mathbf{D_i}(\delta)$ of all the destination nodes that are at distance δ from i, and picks the destination j uniformly at random from these. The regraft height of j – at P_j – will be $h' = 2h(\text{mrca}_{ij}) - h(P_i) - \delta$. See Box 2 for details.

Like SubTreeJump, SubTreeLeap is also not symmetric, hence we need to compute the Hastings ratio $q_\sigma(\tau|\tau')/q_\sigma(\tau'|\tau)$. In order to get back to τ from τ' one would

Algorithm 2: SubTreeLeap transition kernel.

Excluding the root, pick a node i in τ uniformly at random, i.e., with probability $1/(2n-2)$;

Draw a patristic distance δ from the distance kernel $\kappa(\delta|\sigma)$;

Find the set of destination nodes $\mathbf{D_i}(\delta) \leftarrow getDestinations(\tau, i, \delta)$;

Pick a node $j \in \mathbf{D_i}(\delta)$ with probability $Pr(i \to j) = 1/|\mathbf{D_i}(\delta)|$;

Prune the tree at P_i and regraft it at P_j, creating a new tree τ'.

Function *getDestinations*

> **Input :** A tree τ, a node i and a scalar δ.
>
> **Output:** A set $\mathbf{D_i}(\delta) = \{c : d(c,i) \leq \delta\}$.
>
> Determine P_i, S_i and G_i;
>
> Compute $p_i = h(P_i)$;
>
> Compute $h_b = p_i - \delta$;
>
> **if** $h_b > h(i)$ **then**
>
> > Go down the subtree subtended by s_i and find all nodes that intersect at height h_b constructing the set $\mathbf{D_i^{(b)}} = \{d : h(d) \leq h_b < h(P_d)\}$;
>
> **end**
>
> Compute $h_a = h(P_i) + \delta$;
>
> Walk up to the root and find all nodes that intersect at height h_a constructing the set $\mathbf{D_i^{(a)}} = \{d : h(d) \leq h_b < h(P_d)\}$;
>
> **return** $\mathbf{D_i}(\delta) = \mathbf{D_i^{(b)}}(\delta) \cup \mathbf{D_i^{(a)}}(\delta)$

first need to draw the same distance $\delta' = \delta$ from $\kappa(\cdot)$. Then one needs to realise that the original node i is guaranteed to exist in the set of destinations $\mathbf{D_j}$ – but $\mathbf{D_j} \neq \mathbf{D_i}$. Hastings ratio is then

$$\begin{aligned} \frac{q_\sigma(\tau|\tau')}{q_\sigma(\tau'|\tau)} &= \frac{1}{|\mathbf{D_j}(\delta')|} \Big/ \frac{1}{|\mathbf{D_i}(\delta)|}, \\ &= \frac{|\mathbf{D_i}(\delta)|}{|\mathbf{D_j}(\delta')|} \end{aligned} \tag{2.3}$$

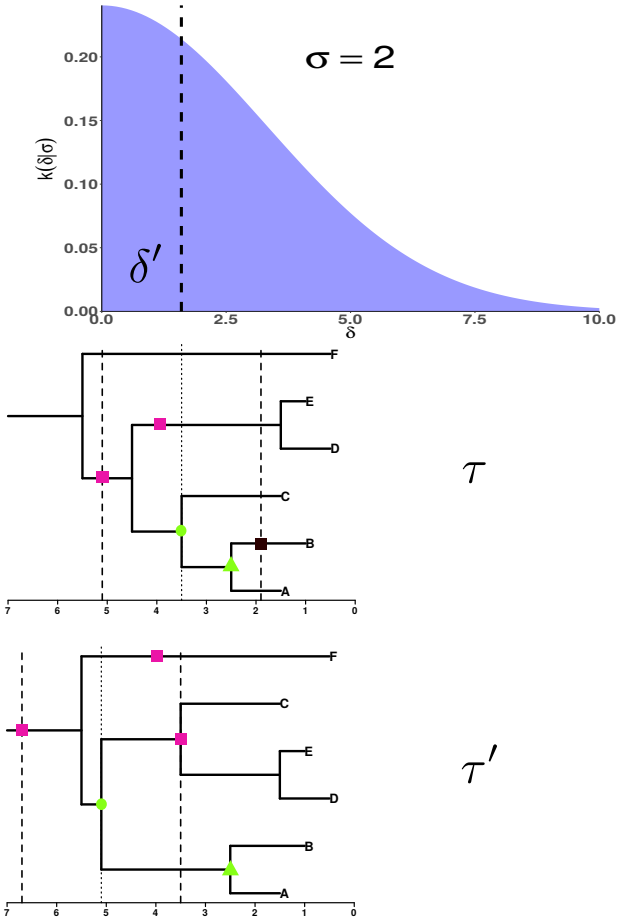


Figure 2.2: Schematic representation of a SubTreeLeap proposal. SubtreeLeap operation with size $\delta' = 1.6$ on τ to obtain the proposal phylogeny τ'. Target node represented by the green triangle and the height of its parent (green circle) marked by a dotted line; valid destination nodes marked with pink squares, invalid (prohibited) destinations are marked with black squares. Notice that (a) there is always a destination node above the root and (b) the Hastings ratio would be $\frac{|\mathbf{D_i}(\delta)|}{|\mathbf{D_j}(\delta')|} = 2/3$. See text for details.

One needs to draw the exact same distance $\delta' = \delta$ to be able to get the original node in the destination set and hence produce τ from τ'. This means the densities

$\kappa(\delta'|\sigma)$ and $\kappa(\delta|\sigma)$ – independently of the choice of distance kernel[6] – cancel out in the Hastings ratio leaving only the discrete component to be computed.

This construction is complicated and deserves a bit more consideration. Let $m_i = \min(g_i, q_i)$ and consider j as a virtual destination node. Depending on the magnitude of the distance δ, a range of rearrangements is possible:

A) Complete slide move (CSM): if $\delta < p_i$ and $\delta < m_i$, the regraft point will lie either on the branch subtending P_i or the one subtended by P_i;

B) Partial slide move (PSM): if $\delta < p_i$ but $\delta > m_i$, there is only one destination – above P_i – because moves that extend p_i below are forbidden due to the restrictions imposed by fixed sampling times.

C) Same subtree topological move (SSTM): if $\delta > p_i$ and/or $\delta > m_i$ but $\delta < h(\rho) - h(P_i)$ a topological rearrangement will occur, but will change only the subtree[7] that contains P_i. Similar to the above, a STM can be either "complete" or "partial", depending on the constraints imposed by m_i;

D) Cross tree (topological) move (CTM): on the other hand if $\delta > h(\rho) - h(P_i)$ the destination set $\boldsymbol{D}_i(\delta)$ will only include destinations on the subtree across i from the root.

For a CSM h' is either $h(P_i) + \delta$ or $h(P_i) - \delta$ with probability $1/2$. The rationale above is valid even when P_i is the root node, meaning the tree can be indefinitely extended above –*i.e.* backwards in time[8]. If $\delta > \max_j d(P_i, j)$ the height h' is bigger than the height of the root (tree height) and there will be no destinations on the opposite subtree. In this case, there is only one destination node, above the root, and the parent of i, P_i, becomes the root (see Figure 2.2). This is in stark contrast

[6] As long as $\kappa(\cdot)$ is strictly positive and unbounded (see Remark 2.2.5).

[7] The root ρ has two children, L and R, and if $P_i \neq \rho$ it is either one of L and R or the child of exactly one of R or L.

[8] This does however require a bit of an abuse of notation, since the root does not have a parent node, but we would have to have $P_\rho = \rho$.

with the default set of operators available in BEAST, where one needs a specific move to change the height of the root. As a trade-off however, `STL` will only change the root height occasionally.

An interesting property of `STL` is that the scaling parameter σ can be set in time units, which makes it easier to tune the parameter to be commensurate with the expected age of the root, for instance. This is particularly helpful when setting the initial value for σ (before chain adaptation) if one has a rough guess of the evolutionary rate. Updating the root height is important because in phylodynamics the height of the root carries information on the time of origin of the most recent common ancestor of the circulating lineages, and sometimes the epidemic (see e.g. Gire et al. (2014)).

2.2.2 The posterior distribution in Bayesian phylogenetics

Before analysing the properties of a `STL` (or `STJ`) Markov chain in phylogenetic space, it would be convenient to study some properties of the target distribution $p(\cdot)$ in order to understand the challenges it poses to MCMC algorithms.

Recall that $\mathbf{\Psi} = \mathbb{F} \times \mathbf{S}$, *i.e.*, phylogenetic space, is an infinite space composed of a finite discrete component of (fully-ranked) *topologies* $\mathbb{F}$, and a continuous space of inter-coalescent intervals $\mathbf{S}$ (branch lengths can also be used). Any analysis of Markov chains on this space needs to take into account the projection of the resulting measure on both spaces, as well as consider their interaction (Gavryushkin and Drummond, 2016). For instance, one can analyse the projection of a random walk on $\mathbf{\Psi}$ by looking at the induced (marginal) random walk on $\mathbb{F}$, represented by the SPR graph (see Chapter 1). Gavryushkin et al. (2018) claim about the MCMC on space of phylogenies: "mixing over these discrete structures is the primary obstruction to MCMC convergence" (pg. 1102). While I do agree with the authors, it is also important to point out that the **interaction** between branch lengths and topology

may also play a role. This is because even if t and $\boldsymbol{b}$ are assumed independent *a priori*[9], they tend to inextrincably correlated *a posteriori*[10]. The main purpose of the candidate-generating mechanisms proposed here is to account for this interaction when proposing new states in the Markov chain. In this section I shall present some observations about the target distribution and its support $\boldsymbol{\Psi}$ in the same spirit as section 6.1 in Dinh et al. (2017), but with the goal of analysing the theoretical properties of phylogenetic transition kernels with special focus on the ones presented here.

Under mild regularity conditions, the likelihood is continuous and smooth up to the boundary of the BHV space; see Dinh et al. (2017) for proofs. This property holds if the phylogenetic prior itself attains the property that it is smooth and continuous up the boundary (Dinh et al., 2017, Assumption 2.3), which I show to be the case for parametric coalescent priors (see Equation 1.6). In general, any function $\mu : \boldsymbol{S} \times \boldsymbol{K} \to [0, \infty]$ such that $\sum_i \int_{\mathbb{R}^{2n-1}} \mu(\boldsymbol{s}, \boldsymbol{k}_i) d\boldsymbol{s} < \infty$ with $\frac{\partial \mu}{\partial \boldsymbol{s}} \in \mathcal{C}^1$ will fulfill these conditions.

Recall that the density under the parametric constant population coalescent is (Drummond, 2002, eq 5.1, pg. 81):

$$\pi_0(\tau|N_e) = \frac{1}{N_e^{n-1}} \exp\left(-\frac{1}{(2N_e)^{2n-2}} \sum_{j=2}^{2n-1} k_j(k_j - 1)(s_j - s_{j-1}) \right). \tag{2.4}$$

It is clear that $\pi_0(\tau|N_e)$ is differentiable with respect $\boldsymbol{s}$. In addition, $\boldsymbol{k}$ does not change inside the orthant, hence the prior is proportional to $\exp(-(s_i - s_{i-1}))$ for any $i \in [2, \ldots, 2n-1]$ and thus a smooth function under the assumption of positive branch lengths.

[9] A dubious assumption, made for computational tractability.

[10] The existence of a particular configuration of branch lengths $\boldsymbol{b}$ is defined only conditional on an underlying topology t.

In order to study the properties of the prior (and by extension the posterior), let us now to make a few observations about phylogenetic space.

Remark 2.2.1. *The mapping* $g : (\mathbb{T}, \boldsymbol{A}) \to \boldsymbol{S}$ *is non-injective surjective.*

Proof. From the method of construction of the intercoalescent intervals (see Chapter 1, section 1.3.1), it is clear that there exists a bijection between $\boldsymbol{A}$ and $\boldsymbol{S}$, i.e., any set of internal node heights $\boldsymbol{a}_I$ can be unambiguously associated with intercoalescent intervals $\boldsymbol{s}$ (and vice-versa), provided one is careful to preserve the indexing. But since the construction of $\boldsymbol{s}$ does not depend on the underlying tree topology, it means there are pairs of points $(t, \boldsymbol{a})$ and $(t^*, \boldsymbol{a})$ such that $g(t, \boldsymbol{a}) = g(t^*, \boldsymbol{a}) = \boldsymbol{s}$. □

Moreover,

Remark 2.2.2. *The mapping* $\psi : (\mathbb{T}, \boldsymbol{A}) \to (\boldsymbol{S}, \boldsymbol{K})$ *is non-injective surjective.*

Proof. For a given tree t with associated node times $\boldsymbol{a}$, the intercoalescent times $\boldsymbol{s}$ and lineages through time $\boldsymbol{k}$ can be thought of as summary statistics. It is possible, however, to have pairs of distinct points $\{t, \boldsymbol{a}\}$ and $\{t^*, \boldsymbol{a}\}$ such that $\psi(\{t, \boldsymbol{a}\}) = \psi(\{t^*, \boldsymbol{a}\}) = \{\boldsymbol{s}, \boldsymbol{k}\}$. To see this, consider the diagram in Figure 2.3, which shows two distinct phylogenies with the same intercoalescent intervals and numbers of lineages, for which the prior measure is the same but the likelihood would not be. □

We then conclude that

Remark 2.2.3. *The parametric prior* $\pi_0(\tau|N_e)$ *is flat in large portions of* $\boldsymbol{\Psi}$.

Proof. Remark 2.2.2 implies that at least one part of the space of interest has the same density under the prior measure for ranked trees (*i.e.* in $\mathbb{T}$). To see that it remains the case on the space of fully-ranked phylogenies, $\mathbb{F}$, notice that as long as the internal node heights coincide, the phylogenies will have the same density

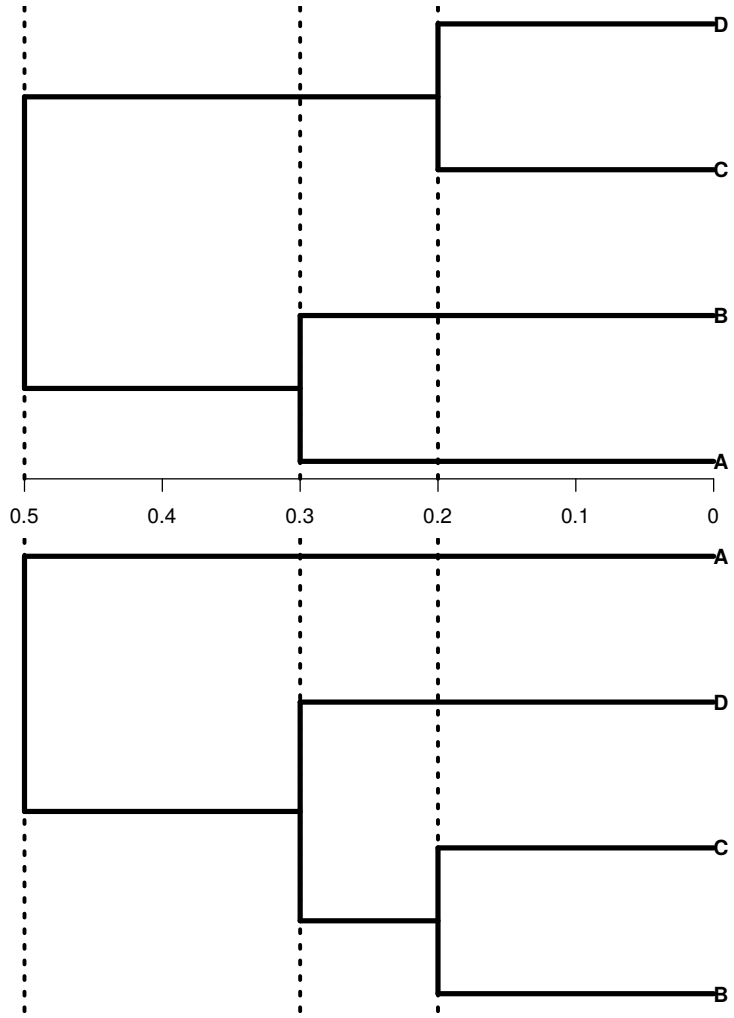


Figure 2.3: Two distinct phylogenies with the same intercoalescent intervals and numbers of lineages. In this example, we would have $\boldsymbol{s} = \{0.2, 0.1, 0.2\}$ and $\boldsymbol{k} = \{2, 3, 4\}$.

under the prior, since the sampling times ($\boldsymbol{a}_L$), which are fixed, induce the same sub-intervals (Minin et al. (2008), Figure 1.). As a consequence, any collection of fully-ranked phylogenies with this configuration would have the same density under the coalescent prior. □

A statistical consequence of Remark 2.2.3 is that the coalescent prior may fail to

regularise any multimodality induced by the likelihood. In a way, the coalescent prior can be seen as an exchangeable prior on $\mathbf{\Psi}$ (Aldous, 1996). Moreover, this prior is uniform on the space of ranked topologies, but induces counter-intuitive priors on individual clades (Pickett and Randle, 2005) complicating interpretations of posterior support for clades[11]. Despite these caveats, the coalescent prior is very common in applications and hence I shall use it as prior measure on $\mathbf{\Psi}$.

2.2.3 Theoretical properties of SubtreeJump and SubtreeLeap

In this section I explore some of the theoretical properties of the candidate-generating mechanisms described here.

Remark 2.2.4. *SubtreeJump does not induce an ergodic Markov chain on* $\mathbf{\Psi}$*.*

Proof. Since `STJ` does not update branch lengths, the result is obvious. However, notice that `STJ` is not irreducible on $\mathbb{F}$ either. To see this, consider a phylogeny $\tau^\star = (t^\star, \boldsymbol{b}^\star) \in \mathbf{\Psi}$ such that for some pair of nodes i, j in $t^\star$ such that $j \neq S_i$, either $h(i) < h(P_j)$ or $h(j) < h(P_i)$. Then the move $i \to j$ is not allowed and hence `STJ` is not irreducible in $\mathbb{F}$. The extreme case is when the condition above holds for *every* pair of nodes, resulting in a perfect ladder tree which is an absorbing state the chain can never leave. □

Since `STL` can change branch lengths, it does not suffer from this limitation. In particular,

Remark 2.2.5. *SubTreeLeap induces an irreducible Markov chain on* $\mathbb{F}$*.*

Proof. First, assume $h(i) > 0 \; \forall i \in V_t$ for any phylogeny $\tau \in \mathbf{\Psi}$ with topology t. Recall that $d_{\text{SPR}}(u, x)$ is the number of SPR operations needed to transform u into

[11] In the interest of fairness, however, it must be said that a uniform distribution over clades is impossible under most prior measures (Steel and Pickett, 2006).

x (section 1.3.1, Chapter 1). Define $\mathcal{N}(x) := \{u \in \mathbb{F} : d_{\text{SPR}}(u, x) = 1\}$ as the *neighbourhood* of $x \in \mathbb{F}$. Ideally, we could show irreducibility if $q_\sigma(y|x) > 0$ for all $x, y \in \mathbb{F}$. However, we cannot make this claim directly, because $q_\sigma(y|x) > 0$ only for $y \in \mathcal{N}(x)$, which is true because there exists $\delta^\star$ such that $P(x \to y|\delta^\star) > 0$ and $\kappa(\delta^\star|\sigma) > 0$ for $\delta^\star > 0$ and $\sigma > 0$ by construction. Since the SPR graph is connected (see Section 1.3.1 in Chapter 1), it follows that any sequence of topologies $\boldsymbol{X} = \{X^{(0)}, X^{(1)}, \ldots, X^{(N)}\}$ where $X^{(i+1)} \in \mathcal{N}(X^{(i)})$ has positive probability under the transition kernel, establishing irreducibility in topological space. □

Theorem 2.2.6. *SubTreeLeap induces an ergodic Markov chain on* $\boldsymbol{\Psi}$ *with respect to* p.

Proof. I will show that `STL` is irreducible and aperiodic, which establishes ergodicity (Meyn and Tweedie, 1993; Roberts et al., 2004; Dinh et al., 2017). First, we need to show irreducibility on $\boldsymbol{\Psi}$ by extending the result of Remark 2.2.5 to include branch lengths. Suppose there exist $\tau, \tau^\star \in \boldsymbol{\Psi}$ such that $\tau^\star$ cannot be reached from τ in finitely many `STL` steps. Following Remark 2.2.5, we may, withouth loss of generality, assume they have the same topology, *i.e.* $t = t^\star$, and that differences between the two phylogenies lie solely in their branch lengths. This would imply that there exist two phylogenies with the same topology that cannot be transformed into one another in finitely many sliding moves, which is clearly false. Note that `STL` has a positive probability of producing a sliding move for any node it picks and all nodes (excluding the root) can be picked for any given `STL` operation.

Now let us show that `STL` is aperiodic. Recall that $\tau = (t, \boldsymbol{b})$ and let $r_\sigma(\tau)$ be the probability that $t = t'$ – conditional on $\boldsymbol{b}$ – and let $A = \{x : r_\sigma(x) > 0\}$. The chain is aperiodic on $\mathbb{F}$ because $p(A) > 0 \forall\, A \subset \boldsymbol{\Psi}$ and $r_\sigma(\tau) > 0 \forall\, \tau$ – according to Tierney (1994) (pg. 1705) this result can be found in Section 2.4 of Nummelin (1984).

Aperiodicity with respect to branch lengths can be shown using a similar argument to the one used above for irreducibility. We will use the concept of Harris

recurrence (Harris, 1956; Chan and Geyer, 1994; Tierney, 1994). First, denote the n-th state of the chain by $X^{(n)}$ and define $\kappa_A = \sup\{n \geq 1 : X^{(n)} \in A\}$ for $A \subset \mathbf{\Psi}$. Following the definitions and results in Roberts and Rosenthal (2006), for our purposes it suffices to show that $\Pr(\kappa_A < \infty | X^{(0)} = x) = 1$ for all $x \in \mathbf{\Phi}$ – since we know that $p(A) > 0$ for all A. Again making use of Remark 2.2.5, we may restrict attention to a family of (sub)sets $A_t = \{x : d_{\mathrm{SPR}}(x, t) = 0\}$ for some $t \in \mathbb{F}$. Now one can use reasoning by contradiction similarly to what was done above to deduce that $\Pr(\kappa_{A_t} = \infty | X^{(0)} = x) = 0$ for $x \in \mathbf{\Phi}$ and for all t. If $d_{\mathrm{SPR}}(x, t) = 0$ we use the sliding move argument above, otherwise we employ Remark 2.2.5 to get us to the case where it is. These arguments establish Harris recurrence of the chain induced by `STL` with respect to branch lengths, completing the proof. □

This establishes the suitability of `STL` for use as the sole phylogenetic transition kernel in a MCMC analysis.

Remark 2.2.7. `STL` *induces a lazy random walk on the SPR graph.*

Since $d_{\mathrm{SPR}}(\tau, \tau')$ is either 0 or 1, it follows that the projection of a `STL` random walk on the SPR graph G_n is a random walk that moves to a neighbouring state with probability $1 - m_\delta$ and stays put with probability m_δ, *i.e.*, an $m_\delta-$lazy random walk. We can compute m_δ from the set of heights $\boldsymbol{H}(\tau)$ and a fixed distance δ by realising that for an `STL` move to *not* result in a topological change, we need to pick a node i such that only a sliding move (SM) is possible. I claim we can write m_δ as follows

$$m_\delta := P(d_{\mathrm{SPR}}(\tau, \tau') = 0) \propto \sum_{j \in V_t} P(\mathrm{SM} | j, \delta), \tag{2.5}$$

$$= \frac{1}{2n-2} \left(1 + \frac{\mathbb{I}(p_L > \delta) + \mathbb{I}(p_R > \delta)}{2} + \right. \tag{2.6}$$

$$\left. \sum_{i \in \boldsymbol{R}(\tau)} \left[1 \wedge \left(\frac{1}{2} \left\{ \mathbb{I}(q_i > \delta) \times \mathbb{I}(p_i > \delta) \right\} + \mathbb{I}(g_i > \delta) \right) \right] \right), \tag{2.7}$$

where p_i, q_i and g_i are as before and $\boldsymbol{R}(t)$ is the set of all nodes of t excluding the root and its two children, L and R. Notice $m_\delta = 1$ when $\delta < \min(\boldsymbol{B}(\tau))$ and

$m_\delta = 1/(2n-2)$ when $\delta > \max(\boldsymbol{B}(\tau))$. It is important however to point out that m_δ depends on τ – and as such should really be written as m_δ^τ – which makes the lazy random walk probability state-dependent.

Correctness

As illustrated in Holder et al. (2005), who show an error in the Hastings ratio of a popular phylogenetic transition kernel, implementing valid MCMC samplers for phylogenetics can be tricky. Incorrect transition kernels can lead to wrong inferences by converging to the wrong target (posterior) distribution or not converging at all. Since phylogenetic space is non-standard, it poses special difficulties to ascertaining the correctness of MCMC implementations, due to its sheer size and the difficulty in obtaining analytical results against which samples can be compared. While some authors opt for two independent implementations, usually in different programming languages – e.g. Drummond et al. (2002) and Dinh et al. (2017) – others choose to validate their samplers by comparing results with other known samplers or theoretical results (Höhna et al., 2008). In this chapter I shall take the latter approach, which I describe in more detail below.

Specifically in the case of phylogenetics, we need to ascertain whether both topologies and branch lengths are sampled correctly. Suppose MCMC is used to approximate the posterior probabilities $P_i = p(T_i|\boldsymbol{D})$, $i = 1, 2, \ldots, F_n$. If $\boldsymbol{X} = \{X^{(0)}, X^{(1)}, \ldots, X^{(M)}\}$ is a Markov chain where each $X^{(j)}$ is a phylogeny sampled at the j-th state, one can approximate P_i as:

$$P_i = \frac{1}{M} \sum_{j=0}^{M} \mathbb{I}(X^{(j)}, T_i), \tag{2.8}$$

where $\mathbb{I}(Y, T_i)$ is an indicator function that is 1 if Y and T_i have the same topology[12]

[12]One can say, for instance, that if the rSPR distance between two trees A and B is 0, then $A = B$.

and 0 otherwise. Branch lengths will be dealt with in a different way (see below). Of course, the bigger $S(n)$, the larger M will have to be in order to obtain good estimates. Since our goal is to assess correctness, it will be convenient to assume that the only parameter of interest is the phylogeny τ (Lakner et al., 2008).

Comparison with samples from the prior

As a baseline for assessing correctness of a transition kernel, one should determine whether its induced Markov chain can accurately sample from the prior. There are several aspects of a MCMC sample that can be analysed with respect to their theoretical expectations of both their continuous and discrete components.

First, we would be interested in determining whether our kernel allows accurately sampling from the – discrete projection of – prior distribution $\boldsymbol{R} = \{R_1, R_2, \ldots, R_{S(n)}\}$ of trees (topologies). In practice, a good estimate of $\boldsymbol{R}$ can be obtained by simulating a large number K of phylogenies from the coalescent prior distribution and calculating the true tree probabilities as described in equation (2.8). To assess correctness, in particular, one can sample then run MCMC for a suitably large number M of iterations, calculate empirical frequencies $\boldsymbol{F} = \{F_1, F_2, \ldots, F_{S(n)}\}$ in the same fashion and then compare $\boldsymbol{F}$ and $\boldsymbol{R}$. If the sampler is correct, these distributions should match each other very closely. One can define an error measure Δ

$$\Delta := \max_{1 \leq i \leq S(n)} \frac{|F_i - R_i|}{R_i},$$

usually called the *maximum relative deviation*.

As the dimensionality of the posterior distribution grows, it becomes progressively harder to accurately sample the distribution of trees, even in the absence of data. Hence, an approach routinely used in practice is look at the distribution of clades instead. Recall that for $n \geq 3$ taxa there are $A(n) = |\boldsymbol{C}| = 2^{n-1} - 1$ possible clades. As $n \to \infty$, $A(n)/F(n) \to 0$, making tracking clades instead of trees an attractive

alternative when dealing with larger data sets commonly encountered in practice (n in the lower hundreds).

Comparing clade distributions can be done analogously to comparing tree (topology) distributions – as exemplified in Section 3.2.4 Chapter 3, the probabilities under the prior can be computed exactly. In particular one can define a similar error measure (Höhna et al., 2008):

$$\delta := \max_{1 \leq i \leq A(n)} \frac{|F_i^c - R_i^c|}{R_i^c},$$

where $\boldsymbol{F}^c$ and $\boldsymbol{R}^c$ are the true (theoretical) and observed probabilities as before.

Coalescent times

In addition to looking at topologies, we also need to ensure the distribution of branch lengths is being accurately sampled. To this end, we can look at the distribution of coalescent intervals. Under a constant population size (N_e) coalescent model, the $k-$th coalescent interval is distributed according to an exponential(λ_k), where $\lambda_k = \frac{k(k-1)}{4N_e}$, $k = 1, 2, ..., n-1$. With this in hand, one can then analyse a sample of trees and assess whether the empirical (observed) distribution of coalescent times matches the theoretical distribution. For instance, one can perform a goodness-of-fit test to ascertain whether the distribution sampled *via* MCMC adheres to its theoretical counterpart.

Dealing with data: marginal likelihoods

Höhna et al. (2008) propose another approach to obtain posterior probabilities, which is to calculate marginal likelihoods for every tree topology. Under the assumption that $\boldsymbol{D}$ was generated by a model $M(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector,

the marginal likelihood for tree T_i is

$$l_i = p(\boldsymbol{D}|T_i) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{D}|T_i, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|T_i)d\boldsymbol{\theta}. \tag{2.9}$$

The claim by Höhna et al. (2008) is that $\frac{l_i}{l_j} = \frac{P_i}{P_j} \quad \forall i, j \in \{1, 2, \ldots, S(n)\}$.

However, this only holds when a uniform prior probability distribution on trees is assumed. When different topologies have different prior probabilities, one must multiply the ratio l_i/l_j by $\pi(T_i)/\pi(T_j)$ before comparing ratios. Here I will present results in log space to reduce numerical instability.

I simulated an alignment with $L = 40$ sites using a five taxa phylogeny (contemporaneous tips) under a simple HKY model with Γ-distributed site-rate heterogeneity ($\alpha = 0.05$). I then ran BEAST for 1 billion iterations, producing a sample of 1 million trees. Marginal likelihoods were calculated by running BEAST with each tree fixed and computing the marginal likelihood using the generalised stepping stone (GSS) method described in Baele et al. (2015). For GSS I used 100 steps with 1 million iterations each ($\beta = 0.3$).

Correctness results

Figure 2.4 shows that all tested kernels are able to approximate the true distribution within 5% absolute error, whilst Figure 2.5 shows that clade frequencies could be estimated below 1% absolute error for all kernels. These results suggest that the moves proposed here correctly lead to the target the distribution of interest in terms of tree topologies and clades. An attentive reader will notice we plot 5% bands for the tree comparisons and 1% bands for the clade comparisons. This is because estimation for clade probabilities is much more precise then for tree probabilities. These thresholds are inherently arbitrary and I feel 5% is an acceptable threshold. I draw attention to the fact that the error measures discussed here are **relative**,

whereas other authors have chosen absolute error loss functions (Höhna et al., 2008; Lakner et al., 2008). In practice this means our thresholds are more strict than previously adopted.

Next, I look at the estimates of the coalescent intervals. As shown in Figure 2.6, all tested kernels seem to produce correct approximations to the distribution of coalescent times. Whilst I could do formal goodness-of-fit tests on the obtained distribution against the theoretical distributions, it often suffices to visually inspect the histogram of coalescent times and check the proximity of the central moments.

The last set of analyses pertains to the behaviour of the transition kernels when targeting the posterior distribution. As suggested by Höhna et al. (2008), the frequency of a particular tree in the posterior sample should be proportional to its marginal likelihood. Figure 2.7 shows the the plot of $\log P_i$ against $\log l_i$ for both the default mix and STL, colouring points by their `rspr` distance to the true tree. In Figure 2.8 I plot log posterior probabilities against corrected marginal likelihoods for every pair i, j such that $i < j$[13]. I confirm the claim by Höhna et al. (2008), adding, however, that direct comparison between l_i and P_i is only possible when assuming a uniform distribution on topologies. In our example, the coalescent prior does **not** assign equal probability across topologies[14] and thus one needs to account for the prior probabilities $\boldsymbol{R}$ in order to obtain a linear relationship between the posterior frequency of a tree and its marginal likelihood.

One thing to note is that, even on a log scale, there seems to be a small bias in the results presented in Figure 2.8, whereby for small marginal likelihood ratios (i.e., more similar trees) there seems to be an overestimation of the ratio of posterior probabilities and conversely for bigger marginal likelihood ratios we see some underestimation. This might be due to instability in the denominator, a common pitfall of ratio estimation. Also, the estimates obtained with SubTreeLeap

[13]Since the ratio correspondence is symmetrical, it suffices to look at the lower triangular entries in the full comparison matrix.

[14]Note that the coalescent prior places a uniform distribution over *labelled histories*.

also seem to be more noisy, although it remains to be seen whether these differences are relevant. Overall, I believe it is possible to claim that both SubTreeJump and SubTreeLeap are correct and target the correct (posterior) distribution.

2.3 Data

2.3.1 Real-world data sets

I compiled a collection of serially-sampled data sets of fast-evolving RNA viruses of various sizes – in terms of number of taxa/sequences and alignment length– in order to have challenging real-world data sets to test the new kernels on (Table 2.1).

Data sets `Dengue4`, `RSVA` and `YFV` are widely used in phylodynamics as teaching data sets and have been extensively analysed. These data sets also have moderate numbers of sequences, which permit better exploration of available methods by allowing more – and longer – chains to be run. To assess performance in bigger, more realistic data sets I composed three more collections of RNA virus sequence alignments, described below.

To compose `denv2Genome`, I downloaded 2382 full genomes (aligned) from Broad Institute's Dengue Virus Portal (`http://www.broadinstitute.org/annotation/viral/Dengue/`), filtered those from serotype 2 and then subsampled to have at most five samples from each year. This resulted in a data set consisting of 90 full genome from DENV-2 isolates from the Americas, ranging from 1987 to 2007. `denv2Env` was constructed by selecting only the envelope sequence for each of the full genomes. Sequences were downloaded pre-aligned.

`flu` is comprised of Human Influenza H3N2 hemagglutinin (HA) sequences ($\approx$ 1700 bp). I downloaded all human HA sequences with more than 1700 base pairs from the Influenza Research Database (`http://www.fludb.org/brc/home.spg?decorator=`

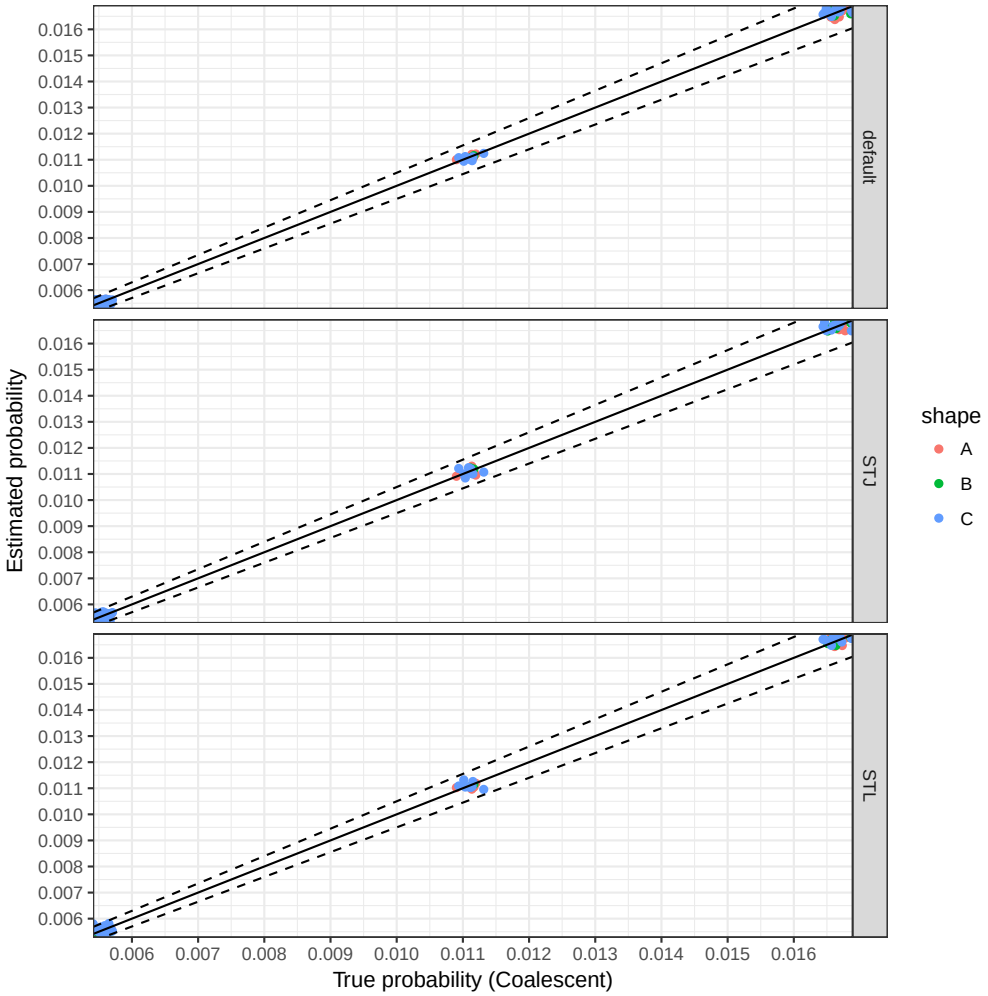


Figure 2.4: Tree probabilities obtained by sampling from the prior with each transition kernel. For each tree kernel, we present the estimated probabilities for each of the 105 possible trees on 5 taxa from a sample of $M = 1,000,000$ trees. On the x-axis, we present the true probabilities, computed from a sample of $K = 100,000$ trees from the coalescent prior by direct simulation. For comparison, probabilities estimated using the default mix of kernels in BEAST v.1.8.4 are provided in the top panel. Solid line shows $x = y$ and the dashed lines show 5% limits. Colours show the three possible tree shapes for 5 taxa.

`influenza`), totalling 8455 sequences, with sampling dates ranging from 1969 to 2014. To make analyses feasible, I downsampled by including at most five sequences

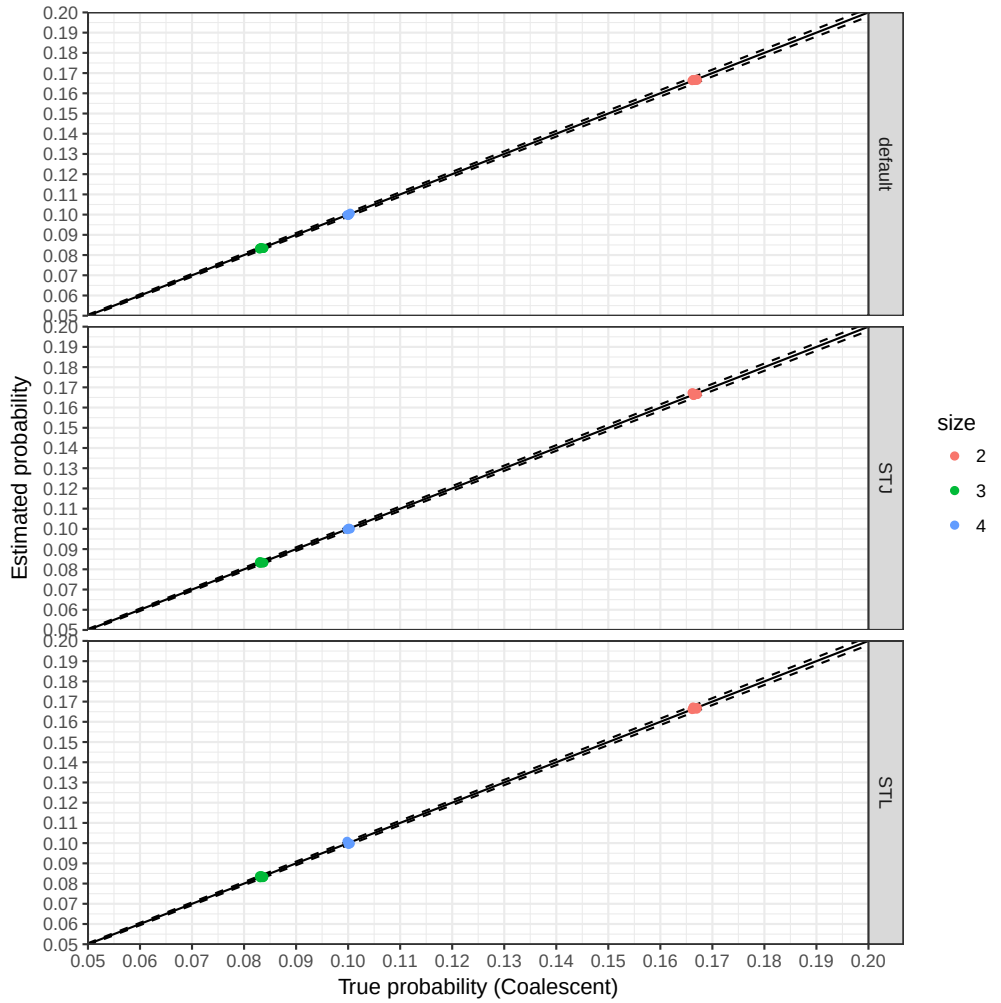
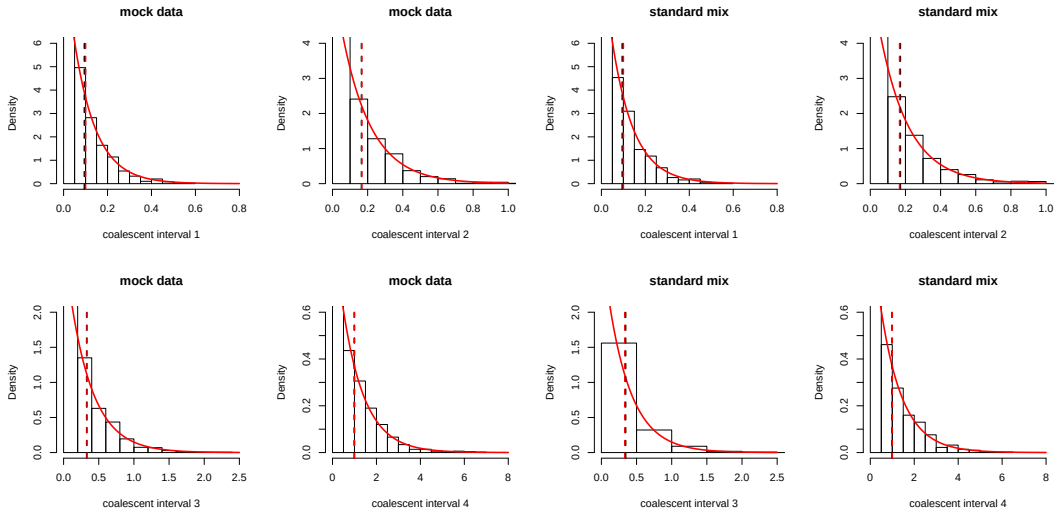
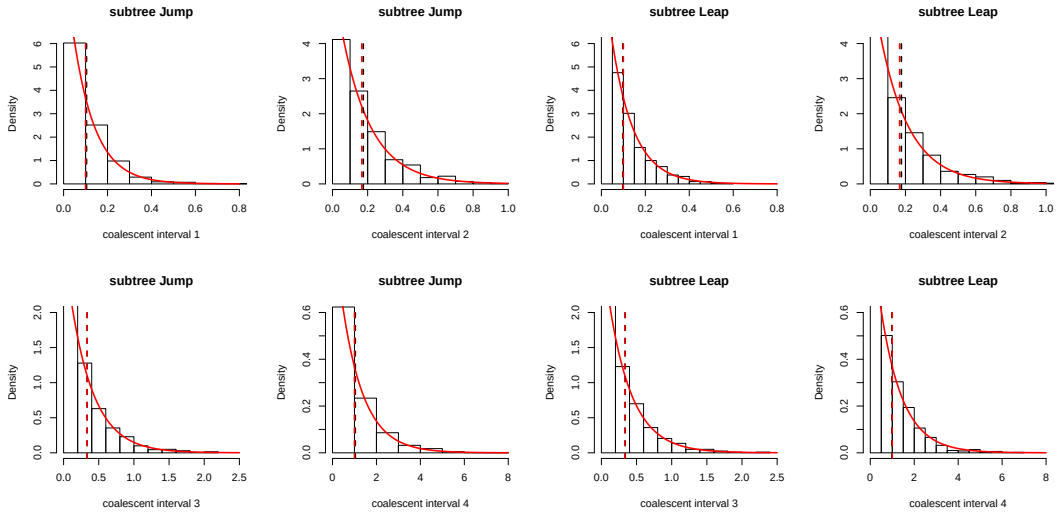


Figure 2.5: Clade probabilities obtained by sampling from the prior with each transition kernel. I used the same sample of 1 million trees to compute clade frequencies to compute the clade frequencies. On the x-axis, I present the true clade probabilities, computed from a the same sample from the prior as before. Probabilities estimated using the default mix of kernels in BEAST v.1.8.4 are again provided in the top panel. Solid line shows $x = y$ and the dashed lines show 1% limits. Colours show the three possible clade sizes for 5 taxa (excluding singletons and the set of all leaves/tips).

from each subsequent year were randomly sampled, resulting in a final data set comprised 225 sequences, which were then aligned by codons using the Geneious software package (`http://www.geneious.com/`).



(a) **Coalescent simulation** (b) **Default operators**



(c) **SubTreeJump** (d) **SubTreeLeap**

Figure 2.6: Coalescent interval distributions obtained by direct simulation, the default (standard) mix of operators in BEAST and our two new kernels. I show the distributions of the four coalescent intervals for $n = 5$ and $N_e = 1000$ obtained by (a) direct simulation from the coalescent process, (b) sampling with the default mix of operators (kernels), (c) SubTreeJump and (d) SubTreeLeap. Red and black vertical dashed lines show the theoretical and estimated means, respectively, and the solid red line shows the theoretical density. To be able to sample with SubTreeJump, we need to combine it with branch length transition kernels, whilst SubTreeLeap can sample on its own.

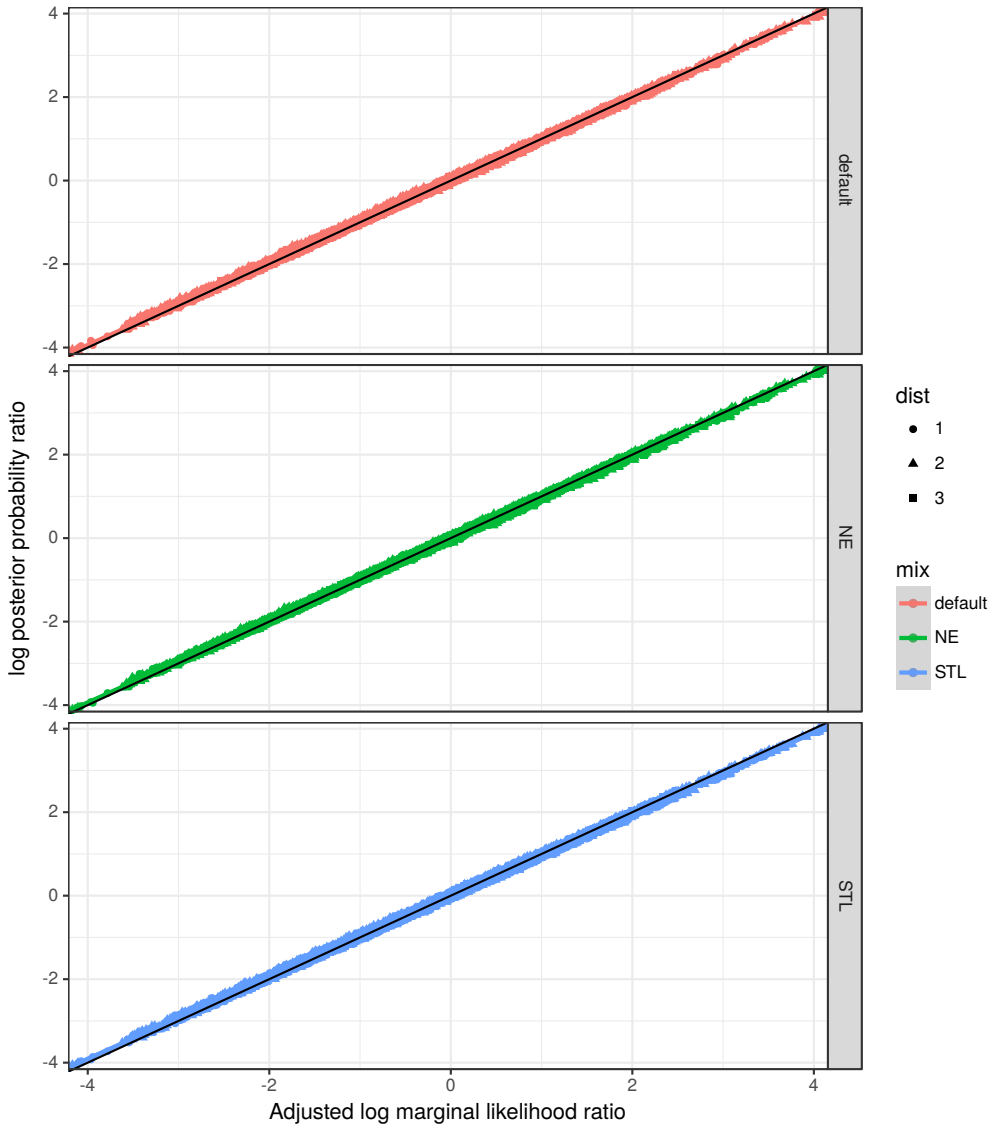


Figure 2.7: Log posterior probabilities versus marginal (log) likelihoods. I plot $\log P_i - \log P_j$ against $(\log l_i - \log l_j) + (\log \pi(T_i) - \log \pi(T_j))$ for the default mix and SubTreeLeap. Points are coloured according to their rspr distance to the true tree. Notice that distance zero means the true tree, used to be simulate the data.

I downloaded all HIV subtype B polymerase (pol) gene sequences from the Los Alamos HIV sequence database (`http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html`) and retained those with known sampling year. This data set comprised 2523 sequences, with sampling years in the period 1983 – 2013.

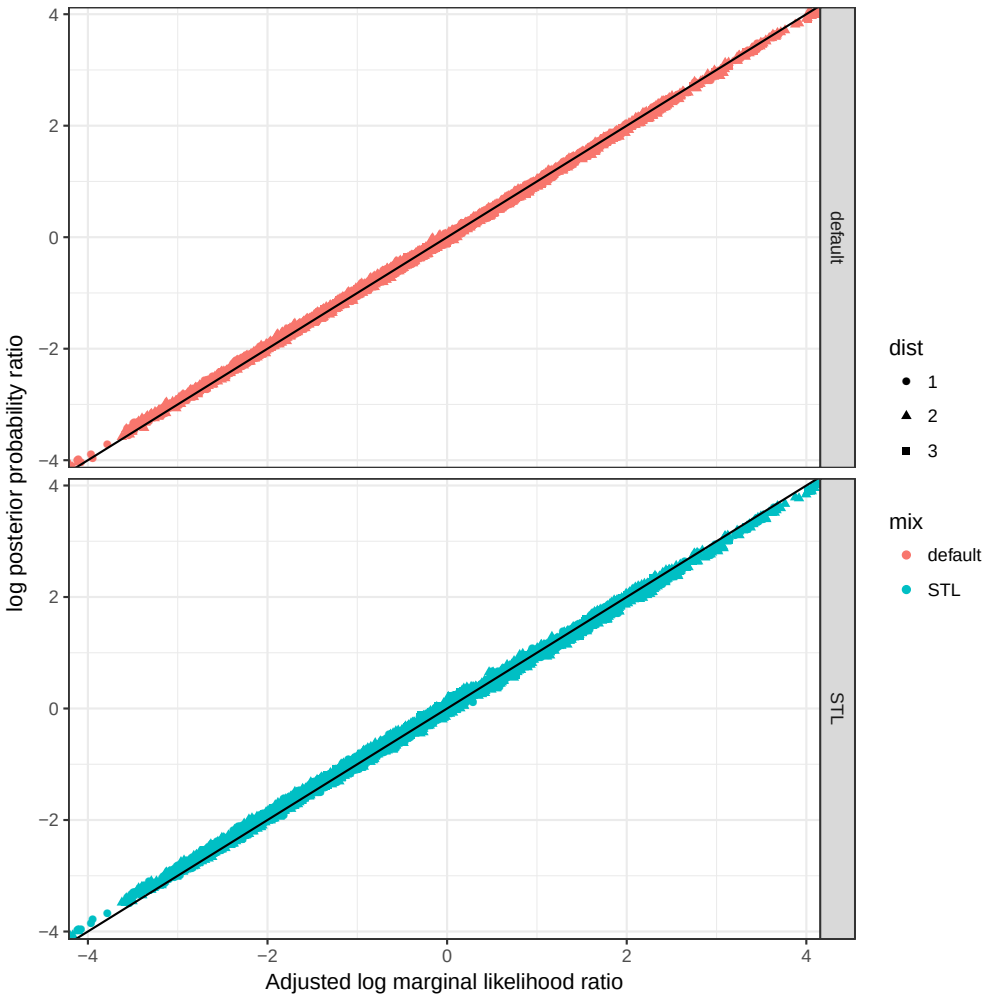


Figure 2.8: Ratios of posterior probabilities of the trees against the ratio of their marginal likelihoods. In this figure I show the (log) ratios as proposed by Höhna et al. (2008). All computations as in Figure 2.7. Solid black line shows $x = y$. Point shapes depict the distance between the pair of trees.

Downsampling was carried out in the same fashion of that for Influenza by keeping the 39 unique sequences from 1983 and sampling from subsequent years, resulting in 187 sequences in the final data set (`HIV`)[15].

Finally, `EBOVa` is one of the largest phylodynamic data sets assembled to date,

[15]The sequences were downloaded pre-aligned and the final alignment was manually checked for inconsistencies.

composed of 1610 serially-sampled full genome sequences of Ebola virus, curated by Dudas et al. (2017). In order to speed up computations without losing realism, I consider a modified version of the data set without the intergenic regions (`EBOVb`). I spend a considerable portion of this chapter discussing ways of improving MCMC performance for data sets `EBOVa` and `EBOVb`, since they represent the type of challenging data set that is quickly becoming the norm in phylodynamics.

2.4 Computational details

At each iteration BEAST picks a transition kernel (operator) at random from the set of transition kernels, with probability $w_i / \sum_i w_i$, where w_i is called the **weight** of operator. An **operator mix** thus is a collection of transition kernels with a certain vector of weights $\boldsymbol{w}$. For the experiments presented in this chapter I have constructed five MCMC schemes employing different candidate-generating mechanisms (operators), described in detail in Table 2.2. I compared the default suite of operators in BEAST (Drummond et al., 2012) to schemes containing `FNPR`, `STJ`, `STL` and a combination of `STJ` and `STL`, which I dubbed `STX`. The idea behind combining `STJ` and `STL` was to help the chain make bigger jumps occasionally, since for $\alpha \geq 0$ `STJ` can lead to bold proposals and hence help with mode-jumping (see Section 2.6.1 however). Notice that when employing either `FNPR` or `STJ` I needed to also include ways of updating branch lengths, since these operators promote only topology changes.

2.4.1 Adaptation scheme

The efficiency of $\bar{\mu_g}$ as an estimator depends crucially on the proposal-generating distribution $Q_\omega(\cdot, \cdot)$, which in turn depends on the indexing parameter ω. In general, ω can be understood as the *width* of the proposal; if ω is too small, consecutive states will be highly correlated, and the chain will not mix well. On the other hand, if

ω is too large, proposed values are likely to have low density under the target and hence get rejected. Ideally, one would want to set ω to an optimal value $\omega^\star$ that maximises the efficiency of chain, as measured by, say, the effective sample size (ESS). In particular, we would like to find $\omega^\star$ such that the acceptance probability α is at its optimal value, $\alpha^\star$. Theoretical analyses of a host of MCMC algorithms for a broad class of target distributions have shown that $\alpha^\star \approx 0.234$ (0.44 for one-dimensional targets) for random walk Metropolis (Roberts et al., 1997, 2001), 0.574 for the Metropolis-adjusted Langevin algorithm (MALA) (Roberts et al., 2001) and 0.651 for Hamiltonian Monte Carlo (HMC) (Beskos et al., 2013).

It is convenient to represent the accept-reject mechanism as a binary-valued process with probability α_ω. In particular, we can write (Andrieu and Thoms, 2008):

$$\bar{\alpha}_\omega := \int\limits_{\mathcal{X}\times\mathcal{X}} \alpha_\omega(x,y)\pi_d(x)q_\omega(x,y)dxdy.$$

Recall that in parallel to the chain $\{Z_i\}$ we have a chain of proposed values $\{Y_i\}$. We can formulate the problem of finding $\bar{\alpha}_\omega = \alpha^\star$ as an stochastic approximation problem, more specifically, we can can write (Andrieu and Thoms, 2008, eq. 17):

$$h(\omega) := E_\omega[H(\omega, Z_0, Y_1, Z_1, \ldots)] = 0,$$

where

$$H(\omega, Z_0, Y_1, Z_1, \ldots) := \min\left[1, \frac{\pi_d(Y_1)q_\omega(Y_1, Z_0)}{\pi_d(Z_0)q_\omega(Z_0, Y_1)}\right] - \alpha^\star.$$

This is the so-called **coerced** acceptance probability case, which is is implemented in BEAST. It is equivalent to finding the zeroes of $h(\omega) = \bar{\alpha}_\omega - \alpha^\star$ (Andrieu and Thoms, 2008).

I shall follow Garthwaite et al. (2016) and assume that the acceptance probability $\bar{\alpha}_\omega$ is a monotonically-decreasing function of the scale parameter. The Robbin-Monro algorithm (Robbins and Monro, 1951) is a popular method for solving the zero-finding problem and consists of creating a positive, non-increasing sequence $\{\omega_i\}$,

$\omega_i : \Omega \times \mathcal{X} \to \Omega$ via an update of the form (Andrieu and Thoms, 2008, eq. 21)

$$\omega_{n+1} = \omega_n + \gamma_{n+1} h(\omega), \tag{2.10}$$

subject to the conditions that $\sum_{n=0}^{\infty} \gamma_n = \infty$ and $\sum_{n=0}^{\infty} \gamma_n^2 < \infty$. One way to attain this is to choose $\gamma_n = \mathcal{O}(n^{-c})$ for $1/2 < c \leq 1$ (Atchadé et al., 2005). In practice, we need to replace $h(\omega)$ with an estimate, for instance

$$\widehat{h}(\omega) := \widehat{\alpha(\omega)}_n - \alpha^\star,$$
$$\widehat{\alpha(\omega)}_n := \sum_{j=C_0}^{n} \alpha_{\omega_j}(Z_j, Y_j),$$

where C_0 is an integer constant chosen so as to avoid transient effects from the initial states of the chain. In BEAST, the Robbins-Monro update is of the form[16]

$$\omega_{n+1} = \omega_n + \frac{1}{f(n)+1} \left(\alpha_{\omega_n}(Z_n, Y_n) - \alpha^\star\right) \tag{2.11}$$

where $f(x) = x$, $f(x) = \log(x)$ or $f(x) = \sqrt{x}$. By default $f(x) = \log(x)$ and $C_0 = \lfloor M/100 \rfloor$. Unfortunately, however, it is not clear to me how $f(x) = \log(x)$ could lead to a valid algorithm, since[17] $\sum_{n=0}^{\infty} \left(\frac{1}{\log(n)+1}\right)^2 = \infty$. In contrast, $\sum_{n=0}^{\infty} \left(\frac{1}{n+1}\right)^2 = \frac{\pi^2}{6}$.

2.4.2 Golden runs

Since we do not know the true posterior distribution for any of the empirical data sets described in Section 2.3.1, I ran very long chains for each data set in order to obtain what we will call "golden runs", which are intended to be a good approximation to the actual target distributions. To obtain these golden runs I ran three independent chains for 10^9 iterations using the default kernels (see above). I

[16] Notice that in BEAST the acceptance rate estimate is **not** smoothed over the chain.

[17] Notice also that $\sum_{n=0}^{\infty} \left(\frac{1}{\sqrt{n}+1}\right)^2 = \infty$.

extracted the last 5,000 phylogenies from each run and (i) obtained a maximum clade credibility (MCC) tree from the resulting 15,000 trees and (ii) computed the 2.5×10^6 pairwise distances between them under various metrics in order to then obtain MDS projections (see Chapter 3). I will call these the "true tree" and "true posterior", respectively. In order to construct target distributions for the empirical data sets I computed the distance from the "true tree" for each of the three golden runs, resulting in 300,000 samples from each distribution (for each metric).

2.4.3 Performance assessment

Lakner et al. (2008) were the first to systematically investigate transition kernel efficiency in MCMC for Bayesian phylogenetics. They investigated the performance of seven kernels on a collection of 10 real-world data sets. To quantify performance, the authors looked at the percentage of converged runs per tested kernel, using clade frequencies relative to a reference (golden) run as a criterion. Time to convergence was also used as performance criterion. Here I will take a similar approach, but with two important distinctions: (a) less reliance on clade frequencies as a convergence criterion and (b) focus on the sampling of continuous parameters that depend on the tree. This is because (a) as the number of taxa grows, it becomes increasingly burdensome to keep track of clades and reliably estimate their frequencies and (b) my ultimate goal is to develop phylogenetic transition kernels that allow quick traversal of phylogenetic space and (indirectly) facilitate sampling of continuous parameters that depend on the phylogeny and hence for properly accommodating uncertainty.

Since each data set presents different difficulties to the sampler(s), different chain lengths are needed to obtain appropriate samples from the posterior. For the performance comparisons I ran 100 independent runs for each operator mix, recording 10,000 samples from the posterior distribution of trees. This was done for `Dengue4` and `denv2Genome` so as to strike a balance between how representative these data sets were of real-world phylodynamic analyses and computational feasibility of running

hundreds of chains. The idea behind this experiment was to explore the performance of the MCMC schemes in more detail, analysing warm-up times and mixing for the two data sets mentioned above.

To study the performance of each our MCMC schemes (operator mixes), I propose to split the problem in two parts: (i) warm-up and (ii) mixing (see Chapter 1 for definitions). The idea is to study (i) how quickly each operator reaches the typical set and (ii) once in a high probability region, how efficiently sampling is done. When analysing simulated data, it is also possible to compute the mean squared error (MSE) for continuous parameters and the effective sample size (ESS) of the distance to true tree (using various metrics) but this possibility will not be explored in the present chapter. For the analyses of empirical data sets, I used the golden runs as ground truth.

To measure warm-up time, one needs to find the iteration i such that $\delta(\frac{1}{i}\sum_{k=1}^{i}\theta^{(k)},\theta) < \epsilon$, for some choice of error function δ and threshold ϵ. Here I will consider a few error functions, specially tailored for phylogenetics. The first error function I propose is the average absolute error in clade frequencies, i.e., the L1 norm between the estimated clade frequencies and their true counterparts – as determined by golden runs. For another global measure of convergence, I propose to find the fraction p_t of the chain at which the distance to the true tree enters the 95% credibility interval of the target as the warm-up time. The rationale behind this is that the sampled trees will initially be more distant from the true tree and once the chain reaches stationarity, samples will remain a certain radius r_d away from the true tree with high probability. This radius, *i.e.* the size of the typical set, depends on the metric (d) used and also on other factors, such as alignment size. The fraction p_t is a measure of how quickly the chain reaches the typical set. A univariate measure of performance could then be to take the maximum value of this fraction across metrics – thus being conservative.

Finally, one can study the warm-up time by considering what fraction p_w the chain

one needs to discard in order to achieve maximum ESS[18] for the continuous parameters. This is done as follows: for a given chain, the **optimal warm-up fraction**, $\hat{p_w}$, is the maximum fraction fraction of the chain one needs to discard in order to obtain the maximum ESS for a given parameter, across all parameters of interest. Here I have chosen the parameters `prior`, `likelihood`, `posterior`, `treeModel.rootHeight`, `treeLength`, `meanRate`, `CP1.kappa`, and `CP3.alpha` because they represent the type of continuous parameter a practitioner would be interested in estimating.

Measuring performance in terms of mixing is an even more delicate issue, because there are several and often incompatible metrics that purport to assess MCMC efficiency. For instance, should one look at wall clock time,*i.e.*, actual time to complete the run, or should we restrict attention to the number of effective samples from the posterior? Here I shall however gloss over some of the nuances in favour of a more direct approach, with well-defined goals. I propose to assess mixing in two main ways: by computing effective sample sizes for continuous parameters (see Chapter 3, Section 3.2.1) and quantifying mixing in phylogenetic space by means of (a) ESS of tree metrics and (b) clade switching (Section 3.2.4).

2.4.4 Analysis of the Ebola virus data set

In order to evaluate the performance of MCMC schemes on a challenging real-world data set, I performed specific analyses on `EBOVa` and `EBOVb`, focusing on a realistic analysis pipeline. To this end I ran three independent runs of 100 million iterations each with the default mix of operators, `STL` and `STX` leading to 9 chains in total. The complete model specification is described in Dudas et al. (2017). I then analysed the resulting MCMC runs in terms of convergence and mixing, as well as performing multi-dimensional scaling under several metrics (see Chapter 3, Figures 3.3 and 3.4).

[18]One can see ESS as a concave function of p_w; too small p_w and ESS will be low due to the transient effect of initial iterations (high autocorrelation); too high p_w and one discards too many samples, bounding ESS above.

2.5 Results

2.5.1 Target distributions

In Figure 2.9 I show the distributions of distance to the true tree for data sets `Dengue4`, `RSVA`, `YFV`, `denv2Env` and `denv2Genome`, obtained using golden runs (see Section 2.4.2). Even in this univariate setting the target distribution can be multimodal and generally non-standard, for most metrics considered. Discrete metrics such as the RF show multiples peaks as expected, since RF for instance can only take values in $\boldsymbol{D}_{\mathrm{RF}} = \{0, 1, \ldots, 2(n-3)\} \in \mathbb{N}$. In Figure 2.9A I show that these distributions however look well-behaved, resembling Poisson distributions (Bryant and Steel, 2009). The distributions for `denv2Env` and `denv2Genome` seem to be translated versions of one another, with the target for `denv2Genome` having lower variance as expected under regularity conditions (more data, less uncertainty about the central tree). Interestingly, the targets for KC metric[19] with $\lambda = 0$ (Figure 2.9B) which reflect topological differences only display very different behaviour, displaying at the same time less granularity and multimodality. While being somewhat obvious, this result does in fact suggest that multiple metrics, even when designed to capture the same features – in this case differences in tree topology – can lead to radically different results (see discussion in Chapter 3 for more).

As expected, focusing on continuous distance metrics that take both topology and branch lengths into account reveals distributions that resemble strictly positive continuous targets encountered routinely in Statistics (e.g. the log-normal distribution). The Steel-Penny (SP) metric seems to lead to target distributions that more closely resemble univariate continuous target distributions; it is unimodal and smooth, albeit with considerable skewness. Notable exceptions are the distributions of SP distances for `denv2Genome`, which presents some clearly defined minor modes (Figure 2.9C, lower left subpanel) and the distributions under the CD metric,

[19] see Chapter 3 for definitions of this and other metrics.

shown in Figure 2.9D. This is interesting because the resulting target for a subset of `denv2Genome`, namely `denv2Env`, does not display these features, providing evidence that multimodality may be strongly data set specific – hence manifesting in some data sets or subsets thereof but not others. The same pattern of the full data set presenting more modes than a subset can be seen in the lower left subpanel of Figure 2.9B, which shows results under the KC metric with $\lambda = 0$ (topology differences only). Distributions of the rooted branch distance (BS) to the true tree show comparable levels of smoothness (Figure 2.9E). In the following sections I will use these target distributions to evaluate the empirical performance of various MCMC schemes (see Section 2.4.3 for methods).

2.5.2 Multimodality in the Ebola 1610 taxa data set

As the results above suggest,a common characteristic of complex discrete-space posterior distributions, particularly Bayesian phylogenetics, is combinatorial multimodality, i.e., multiple peaks composed of atoms (trees) of virtually equivalent posterior density/likelihood separated by valleys of low probability (Lakner et al., 2008; Whidden and Matsen, 2015). In the middle panel of Figure 2.10 it is possible to see that one of the `STL` chains (run 3) gets stuck at a lower density region, which it never leaves. Run 2 (green line) eventually finds the higher density region and samples from it, whereas Run 1 reaches this mode from the start. Also from Figure 2.10 we can see that neither the default set of operators nor `STX` seem to have any problems reaching the higher density region. In particular, `STX` seems to quickly find the typical set – or, more conservatively, the higher mode – and sample from it, while the default kernels take far more iterations to reach the same region. Differences between these modes seem to stem from different topologies being explored, as evidenced by the multi-dimensional scaling (MDS) analysis of the Robinson-Foulds metric (see Figure 3.3 in Chapter 3). Since SubTreeLeap is a an adaptive kernel, the fact that a particular run got stuck at a lower mode could be due to premature tuning of the scaling parameter to a small value, that would in turn make it nearly

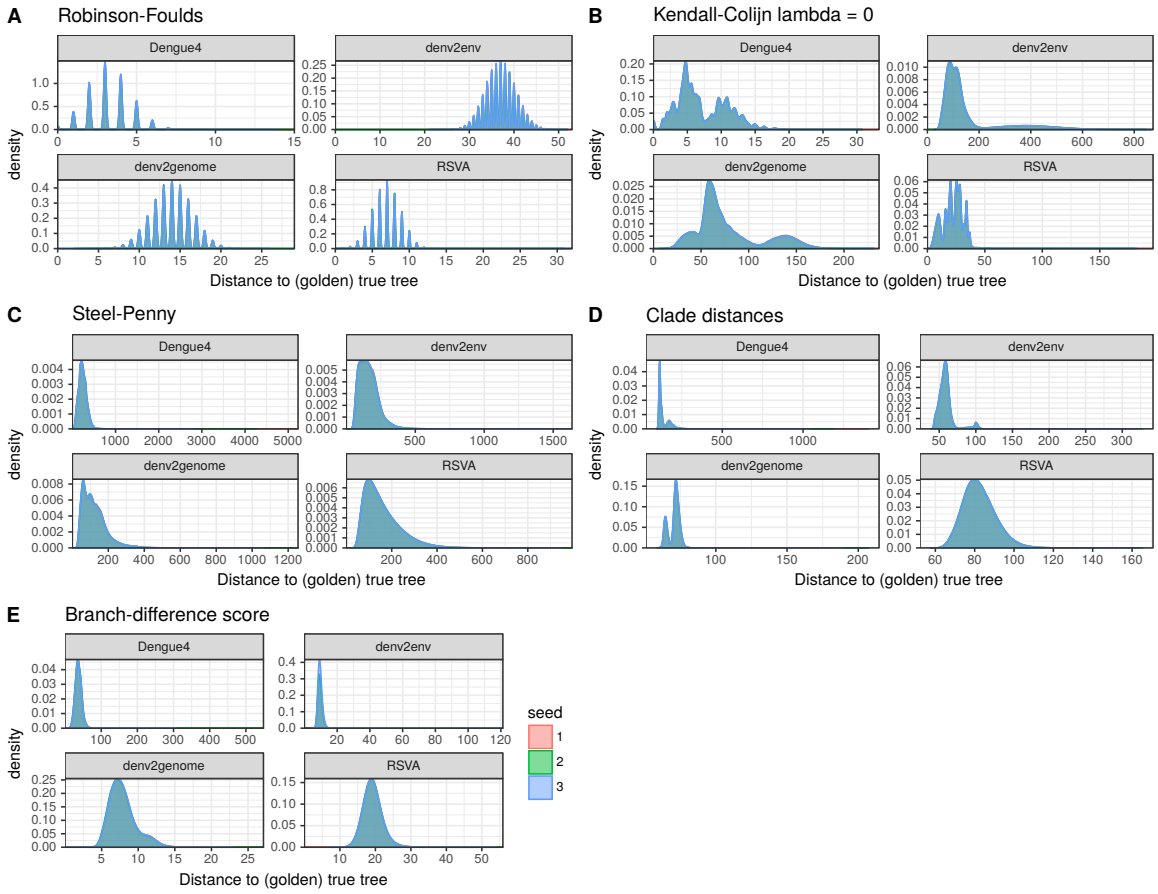


Figure 2.9: Distance to true golden true tree for several data sets and distance metrics (targets). In Panels A and B I show distances that account only for topology differences (Robinson and Foulds (1981) and Kendall and Colijn (2016) (KC) with $\lambda = 0$), whereas panels C, D and E display the distributions of metrics that take branch lengths into account, in the form of the Steel-Penny (Steel and Penny, 1993), clade distance (CD) and rooted branch score (BS), see Chapter 3 for definitions. Colours show the random generating seed (starting value). All kernel density estimates obtained from $100,000$ samples per starting seed (*i.e.* $300,000$ samples in total per data set/metric).

impossible for the chain to transition to a higher mode. Interestingly however, the SubTreeLeap tuning parameter (σ) was similar across all three runs shown in Figure 2.10, with run 1 – which reaches the higher mode from the start – had $\sigma = 0.116$, while runs 2 and 3 tuned to 0.089 and 0.087.

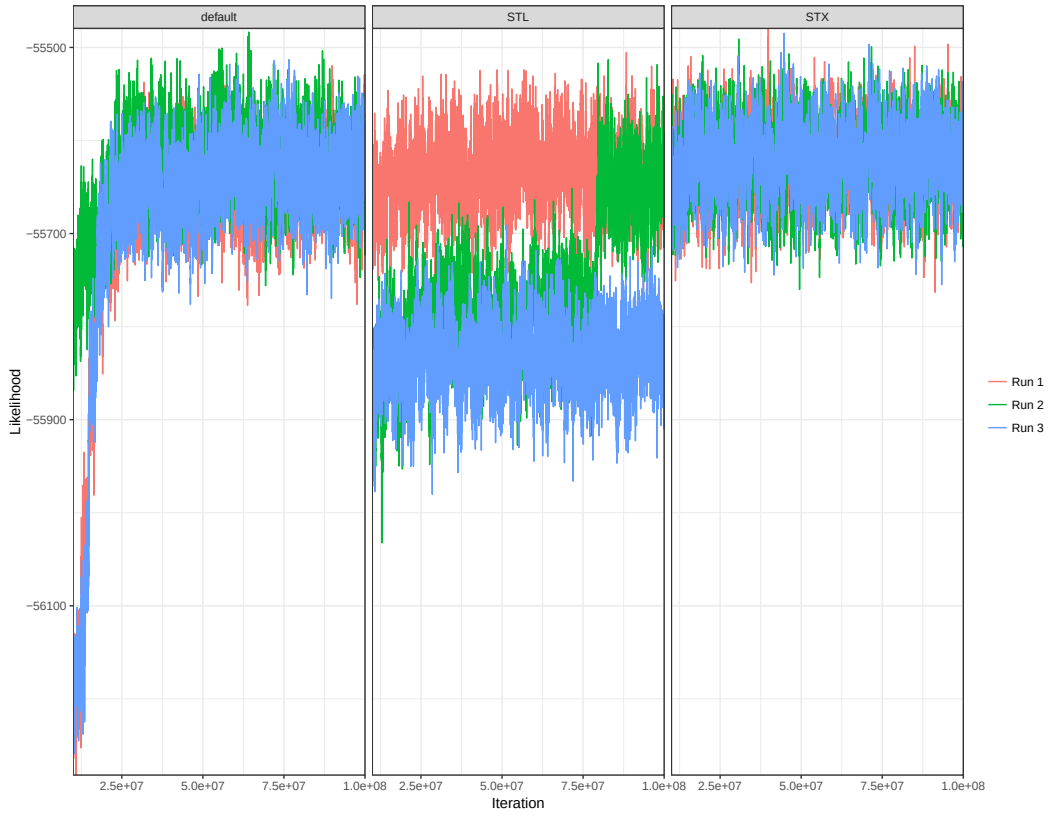


Figure 2.10: Trace plots of the likelihood for the full EBOV 1610 taxa data set. I show the traceplots after 10% of the iterations have been discarded. Colours relate to the seed used in the pseudo random number generator, ensuring each run starts from the same point. These computations were performed with the full data set, EBOVa.

2.5.3 Warm-up, mixing and efficiency

In this section I will discuss the warm-up (burn-in) time, mixing and efficiency of various MCMC schemes considered in this chapter (see Table 2.2). The first quantity I analysed was the fraction p_t of the chain needed to reach the 95% CI of the target distribution, for various metrics (as shown in Figure 2.9). These are shown in Figure 2.11. The first thing to notice is that for most MCMC schemes (operator mixes), only relatively few iterations, under 1% of the total chain length, are required for the chain to start sampling from the bulk of the target distribution. `STL` and

STX showed more difficulty reaching the typical set when compared to the default set of operators. These results are consistent across the two data sets analysed, but the differences in performance are less pronounced for the bigger data set, DENV-2 (90 taxa). Combining STJ and STL into STX improves warm-up quite substantially, even though it does not outperform the default operators. This is in tune with the results for the Ebola 1610 taxa data set, for which STX showed faster convergence compared to STL (see extended discussion).

Moving on to efficiency measures, Figures 2.13 and 2.12 show the average distance attained after 50% of the chain has been discarded as warm-up. The smaller the distance attained, the better the estimates of the mean under a metric d ($\mathbb{E}_{\pi}^{d}$), and the higher efficiency. The results show that STL and STX achieve lower distances and thus higher efficiency of sampling, for both data sets considered. For the larger data set, with 90 taxa, the difference in performance is even larger (Figure 2.13). These patterns remain more or less constant across metrics (e.g. KC, RF or SP), suggesting STX outperforms the other MCMC schemes for both topology and branch length estimation.

The next step was to look at the effective sample size of the distance to the true (golden) tree, as measure of sampling efficiency in that space – under various metrics. This approach is similar to that of Lanfear et al. (2016), who call the ESS of the distance to a focal tree the "pseudo ESS". The main difference is that here I take an MCC tree obtained from three very long (golden) runs as focal tree, instead of the first tree of the chain. From Figure 2.14 we can see that while STL displays better performance than the default scheme, the combination of STJ and STL (STX) has the best performance in terms of the effective sample size of the distance to the true tree[20]. This result seems to be consistent also for simulated data (see Figure B.5 in Appendix B). It seems also that as tree size grows the scaling of STX becomes more important, as evidenced by some of the ESS from the default set of operators being below the common threshold of 200 for some metrics (e.g. the KC metric

[20]Recall these are MCC trees obtained from three separate golden runs.

Figure 2.11: Fraction p_t of the chain needed to hit the typical set for several MCMC schemes and tree metrics. Boxplots show the results of 100 replicates per data set. Vertical tiles show different metrics and horizontal ones show different data sets. The dotted line shows a fraction of 1% of the chain, for comparison. Lower values show faster convergence.

with $\lambda = 0$). If the one-dimensional projections of phylogenetic space provided by the target distributions shown in 2.9 are taken to be faithful representations of the original space then the results above indicate that `STL` and `STX` allow more efficient sampling, not only in terms of distance to the true tree but also in number of effective samples per MCMC iteration.

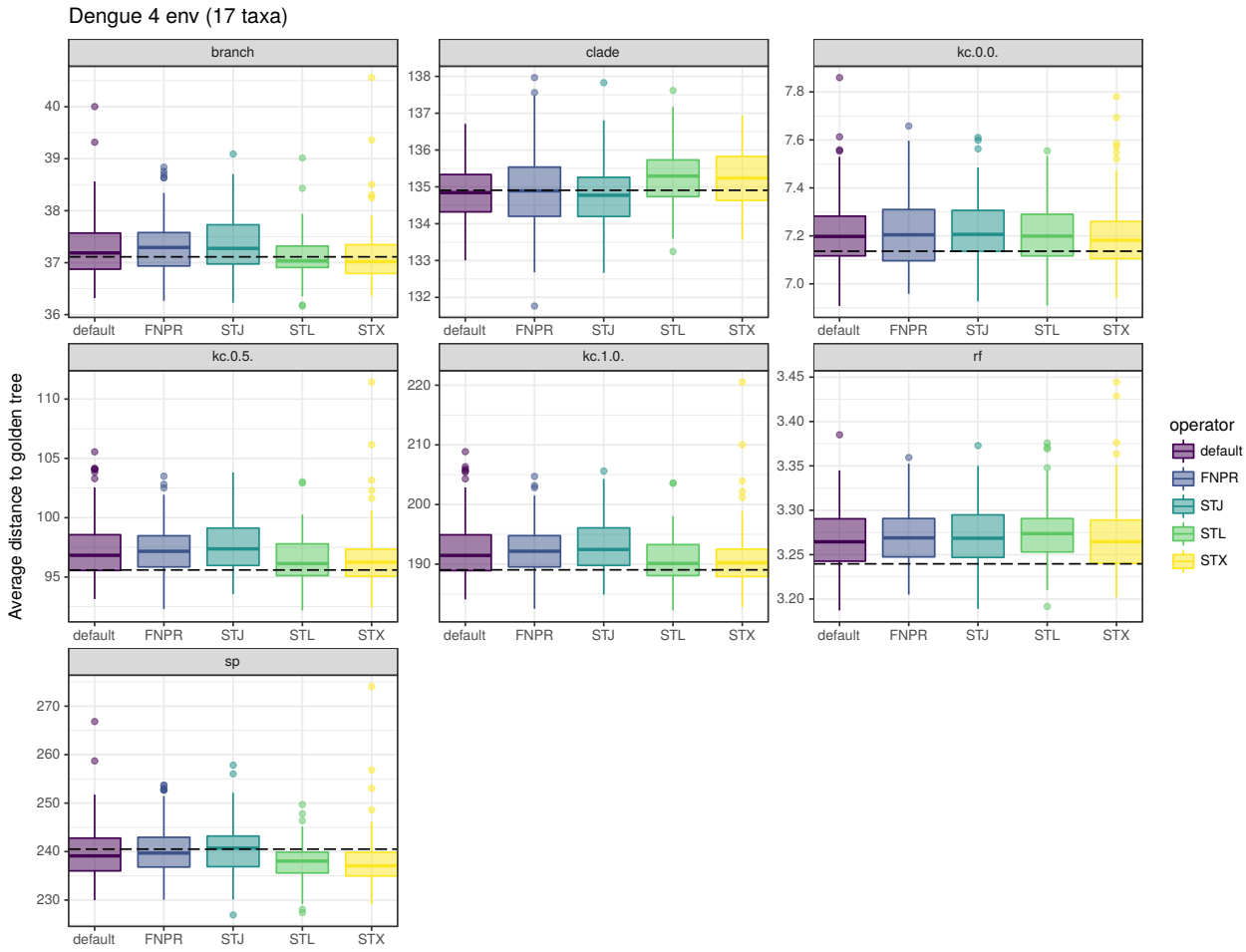


Figure 2.12: Average distance to true golden true tree for several MCMC schemes, Dengue 4 env data set (17 taxa).

Boxplots show the results of 100 replicates per data set. Vertical tiles show different metrics and the dashed lines show the average distances to the true tree computed from the golden runs, i.e. the expectations of the target distributions shown in Figure 2.9.

To further study performance in phylogenetic space, I analysed mixing in clade space. The results in Figure 2.15 show several measures of performance, with panels A and B showing the rate of switching of clade absence/presence indicators (see Chapter 3, section 3.2.4) while panels C and D show the effective sample size

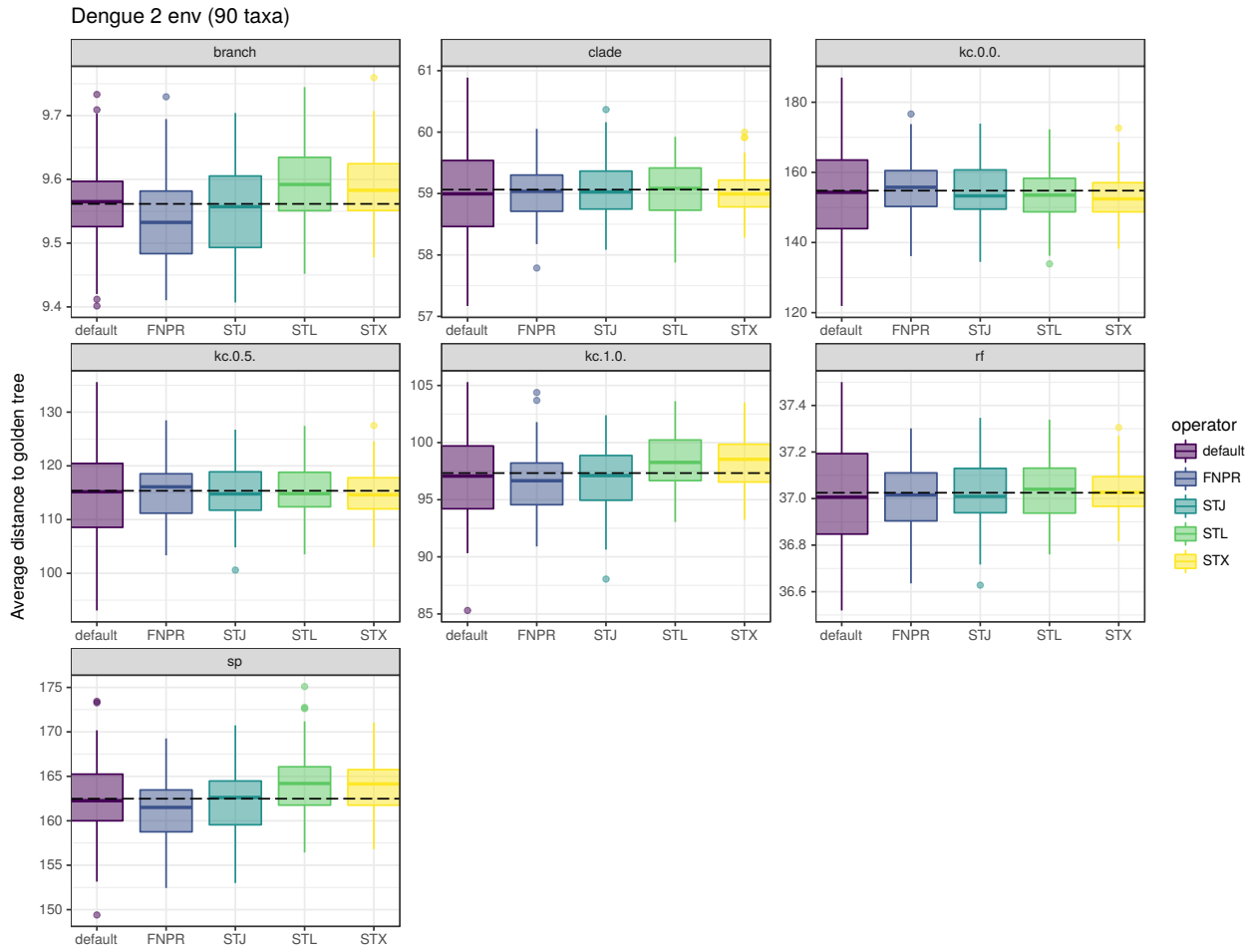


Figure 2.13: Average distance to true golden true tree for several MCMC schemes, Dengue 2 env data set (90 taxa). Boxplots show the results of 100 replicates per data set. Vertical tiles show different metrics and the dashed lines show the average distances to the true tree computed from the golden runs, i.e. the expectations of the target distributions shown in Figure 2.9.

of these clade indicators. In theory, these measures should agree, since they are essentially measuring how quickly the chain explores clade space by visiting each clade proportional to its posterior probability; the faster the switching rate, the higher the ESS (see section 3.2.4 in Chapter 3 for a discussion). And we see that these quantities do agree and show that `STL` and `STX` outperform the other schemes,

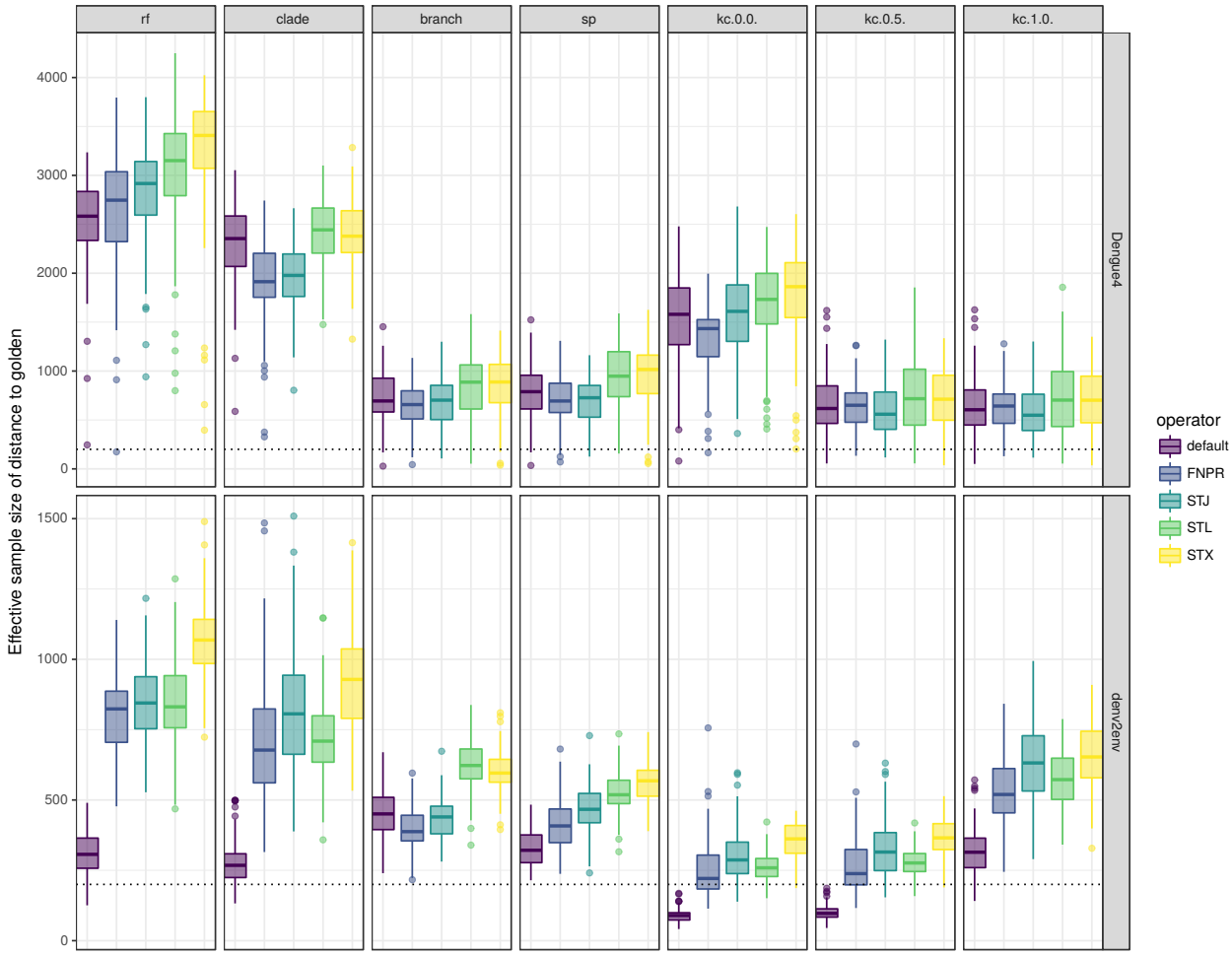


Figure 2.14: Effective sample size (ESS) of the distance to true golden true tree for several MCMC schemes. Boxplots show the results of 100 replicates per data set. Vertical tiles show different metrics and horizontal ones show the two data sets studied in this experiment: DENV 4 *env* (17 taxa) and DENV 2 *env* (90 taxa). The dotted line marks the line ESS = 200.

specially for the larger data set. On the other hand, the plots in Figure 2.15E show that the two new schemes visit slightly fewer clades, but the difference seems to be smaller for the larger data set (see extended discussion).

While performance in phylogenetic space is the focus in this chapter, the ability

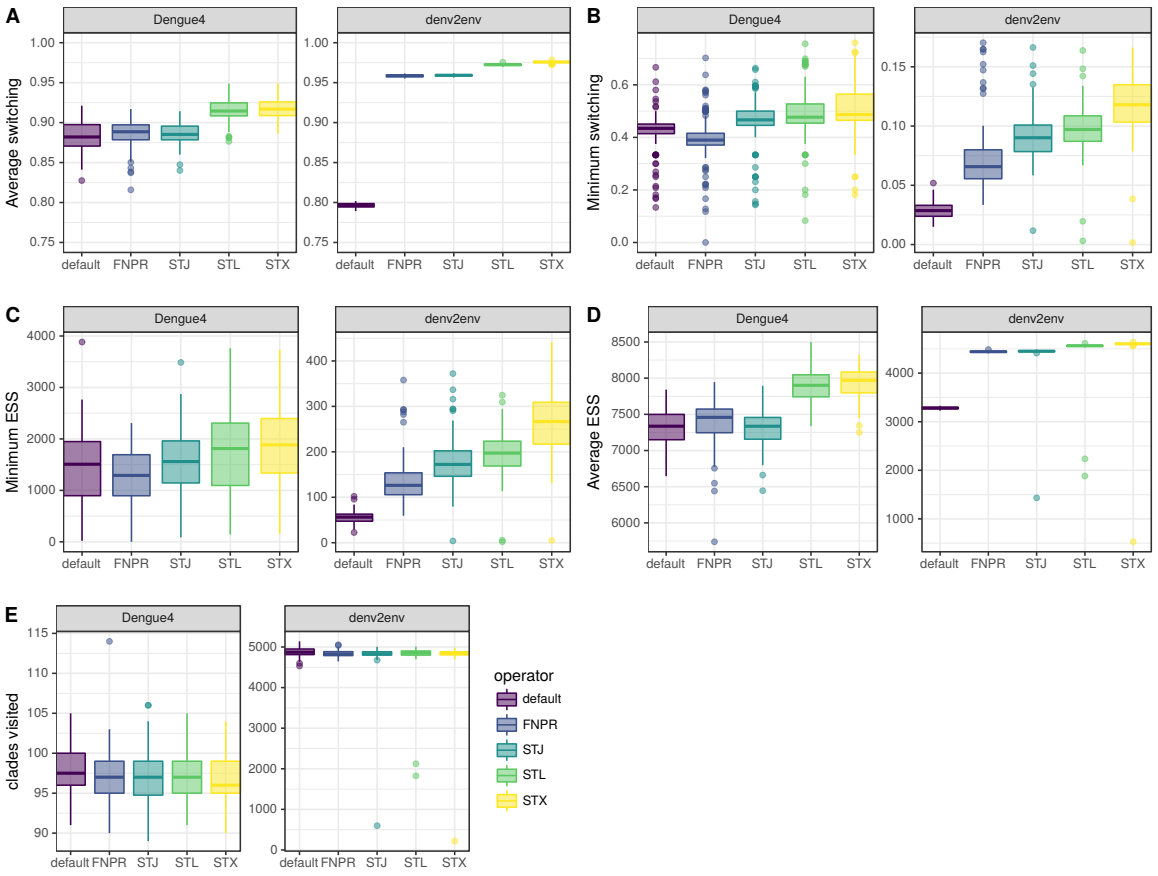


Figure 2.15: Measures of mixing in clade space for several MCMC schemes. Boxplots show the results of 100 replicates per data set. Horizontal ones show the two data sets studied in this experiment: DENV 4 *env* (17 taxa) and DENV 2 *env* (90 taxa). Panel A shows the average clade switching score (see Chapter 3 for definitions), while panel B shows the minimum – across clades – switching score. Panels C and D show the minimum and average ESS for the clade indicators, respectively. Panel E shows the total number of clades visited in a run of 10 (DENV-4, 17 taxa) or 20 (DENV-2, 90 taxa) million iterations with a sample of $10,000$ phylogenies. Please note that y-axes differ between plots.

to quickly traverse the space is also important regarding performance for other parameters that depend on the underlying phylogeny, such as growth and evolutionary

rates. I show that `STL` and `STX` can also facilitate the sampling of important continuous parameters that are dependent on the phylogeny, such as the mean evolutionary rate (`meanRate`). Figure 2.17 shows the effective sample size for two continuous parameters: the mean evolutionary rate and the Gamma heterogeneity parameter α (`CP3.alpha`) for the third codon partition. While the evolutionary rate depends on all branch lengths and rate assignments over the tree and thus constitutes a hard parameter to sample, `CP3.alpha` has very little dependence on the phylogeny (and other parameters in the model) and thus the chain usually converges quite quickly for this parameter. The results show that while all MCMC schemes (operator mixes) perform comparably for `CP3.alpha`, `STL` and `STJ` have improved performance for `meanRate`.

I first looked at the fraction of the chain that, when removed as warm-up (burn-in), maximises the ESS, for a selection of continuous parameters (see Section 2.4.3), taking the maximum fraction across parameters as the measure for a particular run – again, a conservative performance assessment. The results are shown in Figure 2.16 and reveal a slight advantage for `STL` and `STX` for the larger data set (DENV-2, 90 taxa). Overall, however, this performance measure does not indicate substantial superiority of the new proposed schemes over the default.

I then chose two continuous parameters to show the differences in efficiency between MCMC schemes by means of ESS for these parameters. As an example of a parameter for which convergence is usually quick, I chose the Gamma heterogeneity parameter (α) of one of the partitions[21] (`CP3.alpha`) and I chose the mean evolutionary rate, averaged over all the branches of the phylogeny (`meanRate`) as an example of a "hard" parameter, for which convergence and mixing are usually slow. The results, presented in 2.17, show that while for the "easy" parameter (`CP3.alpha`, panel (a)) performance is consistent across MCMC schemes, for the "hard" parameter, which depends more strongly on the phylogeny, the new proposed

[21] Both data sets (DENV-4 and DENV-2) were analysed under three codon partitions.

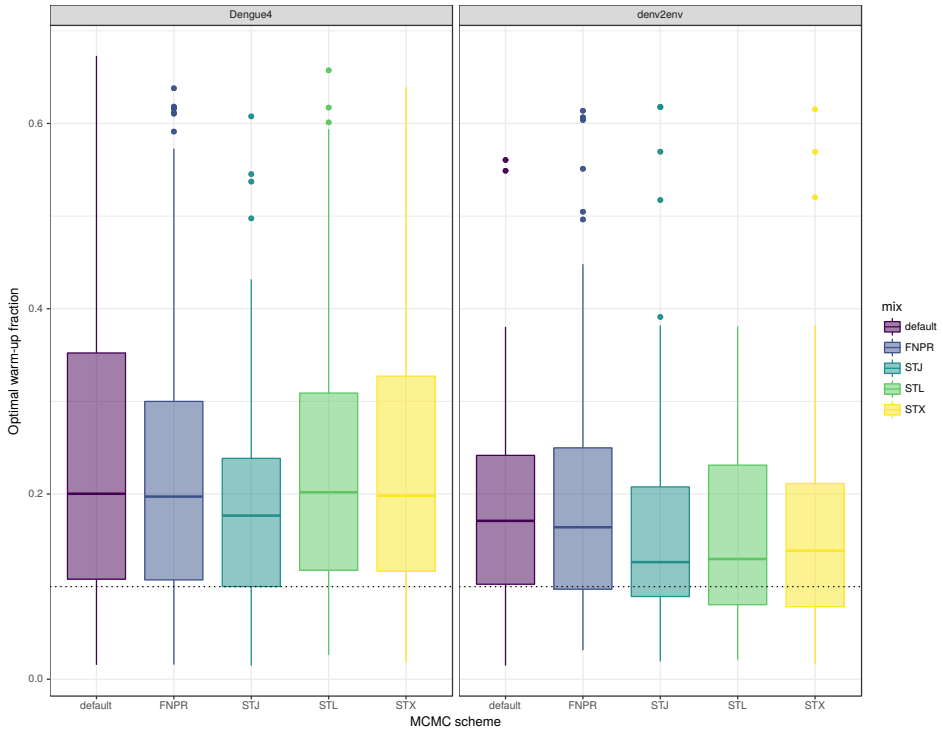


Figure 2.16: Optimal warm-up (burn-in) fraction for several MCMC schemes, continuous parameters. Boxplots show the results of 100 replicates per data set. Vertical tiles show the two data sets studied in this experiment: DENV-4 *env* (17 taxa) and DENV-2 *env* (90 taxa). See text for a description of how the optimal fraction was computed; lower values indicate better performance. The dotted line marks 10%, which is commonly discarded as warm-up (burn-in).

schemes lead to higher ESS. This is specially true for the larger data set (DENV-2, 90 taxa).

Some extra figures showing other results not discussed here can be found in Appendix B.

2.6 Extended discussion

Having already discussed some of the findings as I described them, in this section I shall expand on some aspects, focusing on a more general discussion.

2.6.1 Adaptation issues

While providing performance gains – in terms of mixing – individually, the new proposed schemes also showed considerable difficulty traversing phylogenetic space between modes. This is most apparent in the results in Figure 2.10, where `STL` (middle panel) gets stuck at lower mode, permanently on one chain and for a large portion of the run on another. The results in Figure 2.11 also show that the new schemes have longer warm-up periods.

Another problem we found during development and testing of `STJ` is that for most data sets it was not possible to tune its scale parameter α (see Section 2.2) to attain the desired acceptance probability of 0.234. This is in a way to be expected, because as mentioned above, the set of heights in a any given tree is ultimately discrete, hence not every acceptance probability is attainable. As a workaround, for the EBOV (1610 taxa) analyses I set $\alpha = 0$, i.e., making `STJ` essentially equivalent to `FNPR`, barring differences in computational efficiency due to implementation. This uniform `STJ` when combined with `STL` led to much better performance when compared to the default set of operators.

The results in Figure 2.10 serve as a cautionary tale about the interaction between adaptive kernels and adaptation schemes for multimodal spaces: early adaptation can lead to poor exploration of the space, essentially blinding the algorithm to higher density modes. The increased warm-up times could also be due to poor interaction with the adaptation schedule. Investigating ways of cleverly setting initial values

for the tuning (scaling) parameters could go a long way toward improving warm-up. The findings in Section 2.5.2 suggest that, in terms of the tuning parameter, the difference between a run that never finds a higher mode (run 3) and one that eventually does (run 2) is very small. On the other hand, a bigger size such as the one attained by run 1 offers some protection against getting stuck at a lower mode, as expected.

The findings of Figures 2.15E and 2.11 when taken together, seem to suggest that the inability of `STX` to explore more clades – and hence, potentially, more topologies – in the beginning of the chain could be the reason for larger warm-up times. The adaptive nature of `STL` means that it might sometimes get stuck on a mode because its size has been tuned for optimal sampling of that mode and larger jumps which are necessary for it to traverse the space between modes. I hypothesise that the ability of the default operator mix and `STX` to traverse phylogenetic space is conferred by non-tunable, bold transition kernels such as `WideExchange` in the case of the default mix and `SubtreeJump` in the case of `STX`. These may lead to big tree rearrangements and hence serve as crude mode-jumping kernels. Crude in the sense that they are agnostic about the number and nature of the modes and hence lead to mode-jumping only occasionally and at random. While `STJ` – and presumably also `FNPR` – helps jump modes, ideally we would like one single adaptive transition kernel that could be used alone with comparable or better performance. It is in principle possible to design more guided candidate-generating mechanisms, which exploit the structure of the target distribution to avoid random-walk behaviour (see Höhna and Drummond (2012) and Chapter 6).

2.6.2 Overall perspectives and future research

Overall, the new schemes proposed in this chapter showed better performance, not only in phylogenetic space but also for continuous parameters of interest. In particular, combining `STJ` and `STL` (*i.e.* `STX`) seems to prevent the pathological

behaviour of getting stuck at lower modes while improving performance in a consistent manner. These findings therefore point towards `STX` as a more efficient MCMC scheme, in particular for larger data sets. In most analyses, performance gains were bigger for the larger data set (DENV-2, 90 taxa). Additionally, `STX` considerably outperformed the default scheme for a challenging data set in the form of the EBOV data set which is comprised of 1610 full genomes. Based on the available evidence, I hypothesise that gap in performance between the default scheme (operator mix) and `STL` and `STX` should increase as the size of the parameter space grows – superexponentially – with the number of taxa n. This is because as the parameter space grows, the ability to avoid bold proposals and the resulting rejections becomes more and more important. More technically, I predict that adaptive kernels are progressively more important to combat concentration of measure and the inherent bad scaling of MH methods. More definitive conclusions, however, would depend on specific experiments to investigate scaling of different MCMC schemes as phylogenetic space grows in a controlled manner. In addition, it would be very desirable to investigate the performance of several MCMC schemes for other large, challenging real-world data sets currently being compiled in phylodynamic studies, such as the large ($> 15,000$ genomes) HIV data sets collected by the PANGEA project (Pillay et al., 2015).

A very important question, connected to what was discussed above, is the *optimal scaling* of transition kernels for phylogenies, which might not be optimal for an acceptance probability of 0.234 (Potter and Swendsen, 2015). While the "0.234 result" holds for a broad class of target distributions and MCMC schemes, it is not at all clear to me that the target distributions encountered in phylogenetics attain the necessary regularity conditions for the famous results of Gelman et al. (1996) and Roberts and Rosenthal (1998) to hold. Investigating this issue however would entail substantial development on the theoretical side, since comparatively to what is known for continuous multivariate distributions, we know very little about the geometrical properties of phylogenetic space (Gavryushkin et al., 2018; St. John,

2017; Whidden and Matsen, 2017). I shall touch on a few more issues regarding the application of MH-type algorithms to phylogenetics in Chapter 6.

Another elusive aspect of Bayesian phylogenetics is the correlation structure of the posterior when considering not only the interaction between topology and branch lengths but also dependence of parameters of interest such as R_0 (Stadler et al., 2011) and the evolutionary rate on the phylogeny. Results in Figures 2.11 and 2.16 do not agree: while `STX` has longer warm-up times in phylogenetic space, it seems to lead to faster convergence for continuous parameters. I do not know how to explain these results. It could be that these experiments are not measuring exactly the same phenomenon, since the point from which the chain achieves its the maximum ESS can in theory be quite different from the point where it starts sampling from the typical set. An experiment that characterised the target distribution for continuous parameters using golden runs in the same fashion as what was done in Figure 2.11 could help clarify just how tied together convergence in phylogenetic space and convergence in continuous space are. As a side note, despite giving a precise definition of the typical set in Chapter 1, I have opted to carry out the computations for the warm-up experiments using the (MCMC) time to reach the "bulk" of the distribution in the form of the 95% credibility interval. This is because computing the typical set for continuous distributions would entail extra computations and sensitivity analyses. An interesting topic for future analysis would be to compute the typical set for the targets based on RF metric, which are discrete and for which the entropy is easier to compute, and compare these with the 95% CI.

Finally, in order to see ways for improving our methods, it is of vital importance that we revisit the arbitrary choices made along the way so as to see ways in which generalisation is possible. The first aspect that could be investigated is the balancing between `STJ` and `STL` in terms of weighting: there is no particular reason why the weighting I have used here (Table 2.2) is optimal. It could in fact be the case that one needs to change the relative weighting of the operators as the number of taxa grows, for instance. Another arbitrary choice was made in the

distance kernel distribution used in `STL`. The choice of a half-Gaussian could easily be relaxed, and any continuous distribution with positive unbounded support could be used. Yang and Rodríguez (2013) for instance propose the use of the so-called "bactrian" distributions, which are bimodal distributions for which the "peakness" of the density (*i.e.* the separation between the modes) can be controlled by a tuning parameter m, which could be adapted in much the same way as σ for `STL` or α for `STJ`. These distributions could in theory help the chain get out of lower modes by allowing big jumps occasionally A further arbitrary choice made in the construction of `STL` is to choose the destination node uniformly. There is also no particular reason why the probability of re-grafting P_i at $P_j \in \mathbf{D_i}(\delta)$ should not depend on the actual distance between P_i and P_j. While investigation of the impact these choices on overall performance is of great interest, it would also entail a large amount of experiments to adequately address.

In summary, this chapter provides the following contributions:

- Compiled challenging real-world data sets that better reflect those encountered in practice;
- Proposed an improved framework for checking the correctness of a phylogenetic transition kernel;
- Proposed a framework for evaluating MCMC performance for phylogenetics that circumvents limitations of previous approaches (see also Chapter 3);
- Provided a new adaptive candidate-generating mechanism that updates topology and branch lengths simultaneously.
- Showed improved performance of the proposed candidate-generating mechanisms for challenging real-world data sets.

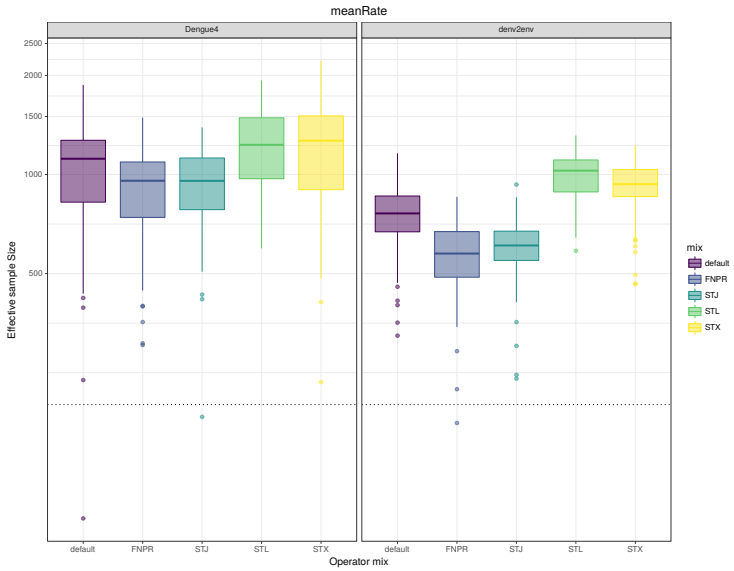
See Chapter 6 for a more general discussion of the relevance of the findings here in the field of Bayesian phylogenetics.

Table 2.1: Collection of serially-sampled data sets used in this study. I report the number of taxa, nucleotide sites and time span (maximum - minimum) of the sequence dates. Denv2 has two versions, one where the alignment consists of full genomes (a) and one with only the *env* gene (b). All data sets were DNA sequence alignments. DENV = Dengue fever virus; RSV = Respiratory Syncytial Virus; YFV = Yellow fever virus; HIV = Human immunodeficiency virus; EBOV = Ebola virus (Makona).

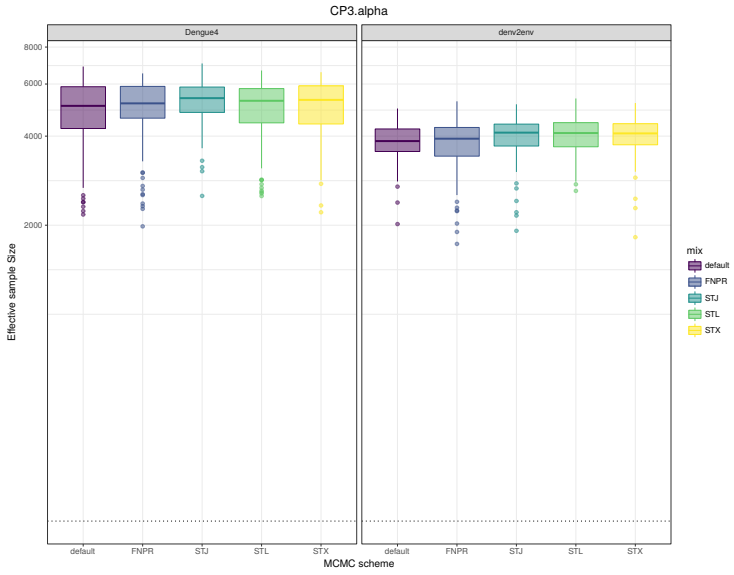
Data set	Type of data	# sequences	# sites	span (years)	Reference
Dengue4	DENV serotype 4 *env* gene	17	1485	38	Lanciotti et al. (1997)
RSVA	RSV subgroup A *G* protein	35	629	46	Zlateva et al. (2004)
YFV	YFV *prM/E* gene	71	654	69	Bryant et al. (2005)
denv2Env/denv2Genome	DENV serotype 2 full genome/*env* gene	90	11851/1441	45	this study
HIV	HIV *pol* gene	187	3012	30	this study
flu	Influenza virus H3N2 HA gene	225	1705	45	this study
EBOVa/b	EBOV full genome	1610	18992/14518	1.6	Dudas et al. (2014)

Table 2.2: Operator mixes used in this study. Each mix was composed of Operator i with weight (w_i). Notice all operator mix (or MCMC scheme) was adjusted so $\sum_i w_i = 69$ to make them comparable with the default in BEAST.

Operator mix	Components (weight)
Classic (`default`)	`subTreSlide` (15), `NarrowExchange` (3), `WideExchange` (3), `wilsonBalding` (3), `rootHeight` (3), `internalNodeHeights` (30)
Fixed-height node prune and regraft (`FNPR`)	`FNPR` (36), `rootHeight` (3), `internalNodeHeights` (30)
SubtreeJump (`STJ`)	`subTreJump` (36), `rootHeight` (3), `internalNodeHeights` (30)
SubtreeLeap (`STL`)	`subTreLeap` (69)
SubtreeLeap and SubtreeJump (`STX`)	`subTreJump` (63), `subTreLeap` (6)



(a) **Mean evolutionary rate**



(b) **Third coding partition α**

Figure 2.17: Effective sample size (ESS) of continuous parameters for several MCMC schemes.. Boxplots show the results of 100 replicates per data set. Horizontal ones show the two data sets studied in this experiment: DENV-4 *env* (17 taxa) and DENV-2 *env* (90 taxa). The dotted line marks the line ESS $= 200$. Panel (a) shows the results for the mean evolutionary rate (meanRate) and panel (b) contains the results for the Gamma heterogeneity parameter α (CP3.alpha).

Chapter 3

Convergence diagnostics for Markov Chain Monte Carlo in Bayesian phylogenetics: the case of time-trees

It is unbelievable that this stubborn darkness, this eternal eclipse, this flaw in geometry, this eternal cloud on virgin truth can be endured.

Farkas (Wolfgang) Bolyai (1775-1856) talks about proving the parallel postulate in a letter to his son János Bolyai (1802-1860).

3.1 Motivation

Markov chain Monte Carlo (MCMC) methods have become a standard tool for approximating complex posterior distributions encountered in Bayesian inference (Robert and Casella, 2011). In phylogenetics, most if not all Bayesian approaches rely on MCMC for approximating the posterior distribution of trees (Li et al., 2000; Suchard et al., 2001; Huelsenbeck et al., 2001). These methods rely on constructing a Markov chain whose stationary distribution is the (target) distribution one wishes to sample from. A fundamental issue is to determine when the chain has reached stationarity and samples are being drawn from the target distribution. Whilst much attention has been given to this issue in the statistical literature, most diagnostic methods assume univariate, continuous parameter spaces. Discrete, high-dimensional parameter spaces such as those encountered in phylogenetics pose additional challenges to development of effective convergence diagnostic tools. In her review of the geometry of tree space, St. John (2017) argues that the power of the tree/phylogeny model "comes from the property that adds the complexity: the vast number of trees to explain different possible evolutionary scenarios" (pg e83). As argued by Drummond and Bouckaert (2015), the complexity of phylogenetic space can be seen as a major reason for the development of specialised software for Bayesian phylogenetics as opposed to the use of common Markov Chain Monte Carlo (MCMC) packages such as Stan (Carpenter et al., 2017) and JAGS (Plummer et al., 2003).

Available methods for diagnosing convergence of MCMC for Bayesian phylogenetics include tracking clade (split) frequencies both within and between chains (Nylander et al., 2008), multi-dimensional scaling of tree distance matrices (Hillis et al., 2005; Matsen, 2006) and network-based clustering (Whidden and Matsen, 2015). These methods are mostly graphical in nature, and only recently have more formal convergence metrics been proposed (Whidden and Matsen, 2015; Lanfear et al., 2016). An important thing to notice is that it is not possible to say with complete

certainty when a Markov chain has converged to its target distribution. Rather, convergence tools are designed to identify failure to converge. As argued by Mossel and Vigoda (2005), when the data do not conform with the model (e.g., come from a mixture of trees rather than a single tree) apparent convergence can be misleading. Cowles and Carlin (1996) and Brooks and Gelman (1998) further reinforce the point that multiple convergence diagnostics need to be employed in order to mitigate the risk of determining convergence when in fact chains have not reached the desired target. Thus, no single method or tool is likely to supersede all the others, as there are cases where one method fails to detect problems but others identify failure to converge. Successful application of convergence detection tools fundamentally depends on combining several metrics/tools in one coherent framework (e.g. the approaches of Nylander et al. (2008) and Lanfear et al. (2016)). An additional issue with currently available methods is that most assume either unrooted trees and/or contemporaneous sequences, limiting their applicability in cases where one deals with time-calibrated phylogenies (time-trees). While Warren et al. (2017) attempt to integrate most of the popular visualisation methods – along with some quantitative indicators – into one framework, their approach is infeasible for large time-calibrated phylogenies.

This chapter is a companion to Chapter 2, where I discuss transition kernels for Metropolis-Hastings MCMC and employ several of the methods described/developed here to assess convergence and performance. My goal here is to expand upon the approach of Warren et al. (2017) and make the necessary adaptations to accommodate time-calibrated trees with hundreds of taxa. In what follows I review some key concepts in Bayesian phylogenetics as well as the state-of-the-art for convergence diagnostics in MCMC applied to Bayesian phylogenetics. I then proceed to discuss the limitations of available methods when dealing with time-calibrated trees and suggest adaptations. My ultimate goal is to investigate a statistically useful representation of tree space and develop an analysis pipeline that can aid practitioners diagnose their MCMC runs when performing phylodynamic analyses for the increasingly larger data sets found in practice.

3.1.1 Tree metrics

A key step in characterising phylogenetic space is constructing valid metrics on it. A tree (phylogeny) metric is a mapping $d_\sigma : \mathbf{\Psi} \times \mathbf{\Psi} \to [0, \infty)$ that satisfies (i) $d_\sigma(u, v) \geq 0$; (ii) $d_\sigma(u, v) = 0 \iff u = v$; (iii) $d_\sigma(u, v) = d_\sigma(v, u)$ and (iv) $d_\sigma(u, w) \leq d_\sigma(u, v) + d_\sigma(v, w)$ for any pair of trees u, v. For convenience, I let σ be an indexing parameter so that we can distinguish between different metrics and study their properties.

A comprehensive review of available tree metrics would be out of the scope of this chapter. Instead, I review a few important metrics that capture different aspects of phylogenetic space. Robinson and Foulds (1981) propose a metric to compare unrooted tree topologies based on a tree operation $\alpha(u, v)$ that removes (contracts) all edges from u that are not present in v, creating $u \wedge v$. The reverse operation α^{-1} in turn adds edges to $u \wedge v$ to create v. Let n_α and $n_{\alpha^{-1}}$ be the numbers of α and α^{-1} operations between u and v. Then the Robinson-Foulds (RF) metric is $d^{\mathrm{RF}}(u, v) = n_\alpha + n_{\alpha^{-1}}$. In other words the RF metric counts (twice) the number of internal branches/edges that differ between two phylogenies. Another important tree metric is the so-called subtree prune-and-regraft (SPR) (Allen and Steel, 2001). Similarly to the above, let β be an operation that picks an edge e in u, prunes the subtree subtended by e and regrafts this subtree at another edge f, creating a new tree v. The SPR metric between u and v counts how many β operations are necessary to turn u into v. SPR is biologically interpretable and can also be used to construct useful representations of phylogenetic space (see Section 3.2.6). While there exist relatively efficient fixed parameter algorithms, computing the SPR metric is NP-hard. As with the RF metric, SPR only captures differences in branching order (topology), not branch lengths.

To address this limitation, Kuhner and Felsenstein (1994) proposed to calculate the square root of the sum of the squared differences of the internal branch lengths corresponding to shared bipartitions between u and v. While this adaptation does

allow comparing branch lengths, it includes topological differences only implicitly. As an advantage however is that there are very efficient, linear time algorithms available to compute the KF distance between pairs and lists of trees (Pattengale et al., 2007). In their excellent review of analytical results on tree metrics, Steel and Penny (1993) propose a simple and easy to compute metric that accounts for branch length differences, called the *path length* difference metric, henceforth called Steel-Penny (SP). Let d_{ij}^{τ} be the sum of branch lengths separating tips i and j in phylogeny τ. The SP metric computes the squared differences of path lengths by two trees x and y as $d_{\mathrm{SP}}(x,y) = \sqrt{\sum_{1\leq i\leq j\leq n}(d_{ij}^{x} - d_{ij}^{y})^2}$.

Kendall and Colijn (2016) proposed a new metric that can be thought of as a compromise in terms of applicability and computational complexity and that allows for the comparison of topology and branch lengths at the same time by cleverly encoding the phylogenies. For a phylogeny u, define $m_{i,j} \in V_u$ to be the first node which is simultaneously the parent of i and j, *i.e.* $m_{i,j}$ is the most recent common ancestor (MRCA) of i and j. Then we are prepared to construct the vector with all $n(n-1)/2$ unique pairs of MRCAs, supplemented by n entries with value 1 at the end, $\boldsymbol{m}(u) = (m_{1,2}, m_{1,3}, \ldots, 1, 1, \{n \text{ times}\}, 1)$. Let $W_{i,j}$ be the branch length associated with $m_{i,j}$ and define l_i to be the root-to-tip path length for tip i in u. Then we can define[1] $\boldsymbol{W}(u) = (W_{1,2}, W_{1,3}, \ldots, l_1, \ldots, l_n)$. For $0 \leq \lambda \leq 1$ we can combine topology and branch length and encode the phylogeny u *via* the vector $\boldsymbol{\eta}_{\lambda}(u) = (1-\lambda)\boldsymbol{m}(u) + \lambda \boldsymbol{W}(u)$. Finally, we are prepared to define the Kendall-Colijn (KC) metric between two phylogenies u and v as $d_{\lambda}^{\mathrm{KC}}(u,v) = ||\boldsymbol{\eta}_{\lambda}(u) - \boldsymbol{\eta}_{\lambda}(v)||$, where $||\cdot||$ is the the Euclidean or L^2 norm. This metric combines topology and branch lengths, is continuous for any $\lambda \in [0,1]$ and is straightforward to compute, specially when compared with the BHV and SPR metrics.

Finally, I describe two other metrics specifically designed for time-calibrated trees. The first, which I shall call the clade-difference (CD) metric, takes two phylogenies τ_A and τ_B and asks how many different clades there exist between them and how

[1] Hence $|\boldsymbol{m}(u)| = |\boldsymbol{W}(u)| = n(n+1)/2$.

different their heights are. More formally, let $\boldsymbol{C}(x)$ be the set of clades of phylogeny x. Then the CD metric can be defined as follows: for each clade $c_i \in \boldsymbol{C}(\tau_A)$, compute the height of the MRCA between the constituents of c_i in $\boldsymbol{C}(\tau_B)$, m_i and the corresponding clade in τ_A. Repeating the same procedure for τ_B we can then compute:

$$d_{\mathrm{CD}}(\tau_A, \tau_B) = \sqrt{\sum_i \left[(h(c_i) - h(m_i))^2\right] + \sum_j \left[(h(c_j) - h(m_j))^2\right]}.$$

The idea behind CD is to quantify big topological differences such that if a clade moves across the tree, the metric will penalise both the difference between the height of that node and the root in τ_A and back down to the height of the corresponding node in τ_B. The second metric is the rooted branch score (BS) proposed by Heled and Drummond (2010) which computes the Euclidean distance in branch lengths between shared clades. Let $b(\tau, c)$ be the length of the branch subtending c in τ if $c \in \boldsymbol{C}(\tau)$ and 0 otherwise. Then we are prepared to define

$$d_{\mathrm{BS}}(\tau_A, \tau_B) = \sqrt{\sum_{c \in \boldsymbol{C}(\tau_A) \cup \boldsymbol{C}(\tau_B)} (b(\tau_A, c) - b(\tau_B, c))^2}.$$

Fortunately, efficient implementations do exist for these metrics, e.g. in the JEBL library (`https://github.com/rambaut/jebl2`).

3.2 Convergence of Markov chain Monte Carlo methods

I now move on to review the existing literature on quantifying and visualising MCMC convergence for Bayesian phylogenetics. My main goal is to explore how several approaches capture different aspects of the process under analysis and discuss their shortcomings when dealing with the types of phylogenies I am interested in this thesis: time-calibrated phylogenies, with large numbers of taxa and complex tip sampling structures. Before I proceed, however, it is important to make clear

that one can never positively assert the convergence of a finite Markov chain to a stationary distribution. In contrast, convergence detection/assessment is done in a negative fashion: one *fails to detect lack of convergence* and hence asserts that **the chain appears to have converged**. Another important remark is that **convergence is a global feature**, *i.e.* for high-dimensional models with many parameters, one can only assert that the run appears to have converged if that is true of *all* parameters.

3.2.1 Convergence diagnostics for continuous parameters

In Bayesian phylogenetics, sometimes researchers are interested in gathering inference about what I will hereafter call "continuous parameters"[2]. These include evolutionary and migration rates, Markov evolutionary model parameters (e.g. κ in the HKY model), amongst many others. Fortunately, this kind of parameter is the standard in models used in the statistical literature at large, which means there is a large body of work on how to detect convergence for continuously-defined quantities in MCMC (see Cowles and Carlin (1996) and Mengersen et al. (1999) for reviews).

Let $\boldsymbol{\theta} = \{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}\}$ be a collection of M samples from a single run of MCMC. Since from a Bayesian perspective all computations should be done from the target posterior p (see Chapter 1), we need to make sure that $\boldsymbol{\theta}$ is a sample from that distribution. The first thing to notice is that the samples in $\boldsymbol{\theta}$ are *correlated* and hence do not constitute a proper sample from the target. Hence we need to quantify the amount of *autocorrelation* between the samples, which in turn will give us an estimate of the *effective* number of samples from π contained in $\boldsymbol{\theta}$, *i.e.* for any $n \leq M$ we want to assess the dependence between $\theta^{(n)}$ and $\theta^{(n+1)}$. I will now make these statements precise, following the notation of Geyer (2011) as much as possible, occasionally borrowing from the Stan manual (StanTeam, 2017) as well.

[2]Called "scalar estimands" elsewhere (Gelman et al., 2014). Here, however, I am also interested in vector-valued quantities, such as root probabilities, population sizes, etc.

It is often the case with MCMC that we are interested in quantity $Y = g(\theta)$, where $g(\cdot)$ a real-valued function, but the distribution of Y cannot be derived analytically and hence expectations of the form $\mu = E[g(X)]$ cannot be computed exactly. An estimate of μ can be obtained from $\boldsymbol{\theta}$ as:

$$\hat{\mu}_M = M^{-1} \sum_{i=1}^{M} g(\theta^{(i)}).$$

If the samples in $\boldsymbol{\theta}$ were independent, we could say that $\hat{\mu}_M \approx \text{Normal}(\mu, \sigma^2/M)$ using the central limit theorem (CLT). However, because by construction the samples are correlated, we need to take that into account when estimating the variance of $\hat{\mu}_M$. Let $\gamma_k = \text{cov}(g(\theta^{(i)}), g(\theta^{(i+k)}))$ be the autocovariances and write the variance σ^2 as

$$\sigma^2 = \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k,$$

and the autocovariances can in turn be estimated as

$$\hat{\gamma}_k = M^{-1} \sum_{i=1}^{M-k} [g(\theta^{(i)}) - \hat{\mu}_M][g(\theta^{(i+k)}) - \hat{\mu}_M] \tag{3.1}$$

With a estimate $\hat{\sigma}_M$ of the variance we can then define the **effective sample size (ESS)**:

$$n_{\text{eff}} = M \frac{\sigma^2}{\hat{\sigma}_M}, \tag{3.2}$$

$$= \frac{M}{1 + 2\gamma_0^{-1} \sum_{k=1}^{\infty} \gamma_k}, \tag{3.3}$$

which captures the number of samples in the MCMC sample $\boldsymbol{\theta}$ that are effectively independent and hence can be used for inference[3]. In Bayesian phylogenetics practice, ESS is at the core of most convergence assessment – for continuous parameters –, the common recommendation being that all parameters in a MCMC

[3]In practice the upper limit of the summation in the denominator of Eq (3.3) is substituted by a finite bound K. See Geyer (2011) and references therein for details on the determination of K and the estimation of n_{eff}.

run have $n_{\text{eff}} > 200$. The rationale for this recommendation seems to be that the variability in most quantities of interest is such that a sample size of 200 should allow for precise computation of most functionals. In this chapter I will use the `effectiveSize()` in the **coda** package (Plummer et al., 2006) and the routines in the package Tracer v1.7 (Rambaut et al., 2018) to compute ESSs. Note however that while the ESS is a useful tool for assessing mixing, it is designed to deal with one chain (run) at a time.

With respect to *convergence* to the target distribution, however, ESS may be misleading in the sense that chain stuck in a mode can have high ESS but has not converged to the target distribution because it has not adequately explored all of the modes proportional to their probability. A powerful technique for assessing convergence is to run several parallel chains from overdispersed starting points and determine whether they converge to the same distribution. Suppose K runs of M iterations each are available, *i.e.*, we have $\boldsymbol{\theta}_i, i = 1, 2, \ldots K$. The between sample variance can be written as

$$B = \frac{M}{K-1} \sum_{k=1}^{K} \left(\bar{\theta}_k - \bar{\bar{\theta}} \right)^2, \tag{3.4}$$

where $\bar{\theta}_k = M^{-1} \sum_{i=1}^{M} \theta_k^{(i)}$ and $\bar{\bar{\theta}} = K^{-1} \sum_{k=1}^{K} \bar{\theta}_k$. Now we can define the within variance as

$$W = K^{-1} \sum_{k=1}^{K} s_k^2, \tag{3.5}$$

$$s_k^2 = (M-1)^{-1} \sum_{i=1}^{M} \left(\theta_k^{(i)} - \bar{\theta}_k \right)^2. \tag{3.6}$$

Finally we can define the **potential scale reduction factor** (PSRF) (Gelman and Rubin, 1992):

$$\hat{R} = \sqrt{\frac{(M-1)W + B}{MW}}. \tag{3.7}$$

At convergence, $\hat{R} < 1.1$, providing a univariate measure of convergence across chains (for a given parameter).

Finally, I note that we are usually interested in complex models, where correlation between parameters is frequently a feature of the posterior. I detail the approach of Vats et al. (2015) that attempts to provide an overall measure of convergence by means of a **multivariate effective sample size (mESS)**. The idea is to accommodate posterior dependence between parameters by jointly considering all parameters at once and estimating their variance-covariance matrix. For M samples, the mESS is defined as :

$$n_{\text{eff}}^{m} = M \left(\frac{\det(\mathbf{\Lambda})}{\det(\mathbf{\Sigma})} \right)^{1/p}, \tag{3.8}$$

where p is the number of parameters under analysis and $\det(\mathbf{\Lambda})$ and $\det(\mathbf{\Sigma})$ are the determinants of the sample covariance matrix and the Monte Carlo covariance matrix, respectively. Notice that similarly to the ESS, when samples are independent, $\mathbf{\Lambda} = \mathbf{\Sigma}$ and hence $n_{\text{eff}}^{m} = M$. Routines to compute the mESS are implemented in the R package **mcmcse** (Flegal et al., 2017).

These procedures can also be used to determine the lower bound on the multivariate ESS to achieve a certain level of confidence. If ϵ is the fraction of the variance due to Monte Carlo error and we would like to collect enough samples to have $(1-\alpha)\times 100\%$ confidence, the bound becomes

$$n_{\text{eff}}^{m} \geq \frac{\pi 2^{2/p}}{(p\Gamma(p/2))^{2/p}} \frac{\chi^2_{1-\alpha,p}}{\epsilon^2}, \tag{3.9}$$

where $\Gamma(\cdot)$ is the (analytically continued) Gamma function and $\chi^2_{1-\alpha,p}$ is the appropriate quantile of a chi-squared distribution with p degrees of freedom. For instance, if $p = 1$ and one would like for the Monte Carlo error to be no more than 5% ($\epsilon = 0.05$) of the total variance and would like to have 95% confidence, we would need an ESS of at least 6146. If Monte Carlo error is allowed to be 10% ($\epsilon = 0.10$) the minimum required ESS decreases to 1516, a figure more than seven times higher

than the folkloric recommendation for ESS $\geq$ 200 in phylogenetics[4]. If one adopts this recommendation –*i.e.* fixes ESS = 200 –, keeping $\alpha = 0.05$ gives $\bar{\epsilon} = 0.27$, meaning this recommendation leads to inferences being drawn from samples where nearly 30% of the variance is due to Monte Carlo error.

3.2.2 Convergence in phylogenetic space

I now move on to present the state of the art for convergence metrics specifically designed for phylogenetics. Their limitations are also explicitly discussed.

3.2.3 Clade frequencies

When diagnosing MCMC convergence in phylogenetic space, simply inspecting the traces for, say, the phylogenetic likelihood, might be misleading because two phylogenies with similar likelihoods may not be necessarily close in phylogenetic space. The approach of Nylander et al. (2008) is to analyse clade/split frequencies to this end. The program AWTY (short for "are we there yet?") developed by Nylander et al. (2008) provides graphical facilities for assessing convergence by analysing various aspects of the distribution of sampled clades. I shall describe and discuss these features now.

For $i = 1, 2, \ldots, |\boldsymbol{C}|$, let $\boldsymbol{X}_i = \{X^{(1)}, X^{(2)}, \ldots, X^{(M)}\} \in [0,1]^M$ be a collection of samples from a Markov chain such that $X_i^{(j)} = 1$ if clade i was sampled in the j-th iteration and 0 otherwise. Also, for $s_i = \sum_k X_i^{(k)}$ we call $f_i = s_i/M$ the *frequency* of clade i. By plotting clade frequencies estimated in two independent runs against each other (scatterplot), one can assess whether both chains have converged to similar distributions. Lack of convergence can be detected when points fall away from the identity ($x = y$) line. For a single run, one useful diagnostic is plotting cumulative

[4]See for instance `http://beast.community/analysing_beast_output`.

clade frequencies along the chain. If these trajectories present long-term trends[5], it means clade frequencies have not stabilised, indicating lack of convergence.

When multiple (say K) runs are available, a very common univariate summary associated with clade frequencies is the **average standard deviation of split frequencies (ASDSF)**:

$$\delta_C = (K-1)^{-1/2} \sum_{i=1}^{C} \sqrt{\sum_{k=1}^{K} \left(f_{ik} - K^{-1} \sum_{k=1}^{K} f_{ik} \right)^2} \tag{3.10}$$

where C is the number of clades seen in all runs. ASDSF is employed for instance in the software `Mr Bayes` (Huelsenbeck and Ronquist, 2001), where an ASDSF of less than 0.01 is taken as sign of convergence – and used as a stopping rule.

Absence-presence plots show whether a particular clade was absent or present in the tree sampled at each iteration of the chain. If there are long periods where the clade is either absent or present, this indicates the chain has not mixed well and might not have converged. On the other hand, a trace plot of this kind where the indicator variable frequently switches between 0 and 1 indicates good mixing. This notion of "clade-switching" can be made more precise (see section 3.2.4). It should be noted however that tracking all clades present in a given run becomes exponentially more cumbersome as the number of taxa increases, quickly overwhelming any graphical diagnostic capabilities.

Finally, one can also plot the (phylogenetic) distance within and between runs using any of the metrics described above. The idea is, again, that at convergence these sets of distances should be similar, and I discuss below (Section 3.2.7) how to make this notion mathematically precise following Whidden and Matsen (2015). In short, AWTY provides the following diagnostics: (i) scatterplot of clade frequencies; (ii) plot of cumulative clade frequencies; (iii) absence-presence plots for clades and (iv)

[5] I am not familiar with the original implementation of AWTY, but this could be made precise by for instance applying LOESS-based trend detection methods to the trajectories.

tree distances between and within runs. See Figure 1 in Nylander et al. (2008) for a graphical summary. A modern incarnation of AWTY, RWTY (Warren et al., 2017) seeks to provide modern plotting routines to implement the AWTY framework in the R language. RWTY also integrates other techniques not available from AWTY like multi-dimensional scaling visualisation (see below). The implementation of RWTY is amenable to automation, insofar as it allows easy scripting and report generation using capabilities available for R.

Limitations

Most methods in AWTY (and RWTY) are visual, and hence do not provide precise quantitative measures to guide researchers toward a decision. This in turn also means that these procedures are hard to automate, what hinders their application in automatic phylogenetic pipelines such as `NextStrain` (Hadfield et al., 2017). As an example, consider the clade frequency scatter plots described above: what is an acceptable bound on the deviations in clade frequencies between two independent runs? This question could be tackled with theoretical considerations and/or a careful empirical study, but neither Nylander et al. (2008) nor Warren et al. (2017) offer any insight into the matter.

3.2.4 Clade switching

Here I describe what to the best of my knowledge is a new metric for quantifying mixing in clade space. Let $\boldsymbol{X}_i = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\} \in [0,1]^n$ be a collection of samples from a Markov chain such that $X_i^{(j)} = 1$ if clade i was sampled in the j-th iteration and 0 otherwise. Also, for $s_i = \sum_k X_i^{(k)}$ we call $f_i = s_i/n$ the *frequency* of clade i. For $m_i = \min(n - s_i, s_i)$, it can be shown that the maximum number of transitions that can be observed from $\boldsymbol{X}_i$ is either $J_i = 2m_i$ [6].

[6] Technically, J_i depends on the first state $X_i^{(1)}$. Suppose w.l.o.g. that $m_i = s_i$. Then $J_i = 2m_i - 1$ if $X_i^{(1)} = 1$ and $J_i = 2m_i$ otherwise.

When the chain is mixing well, *i.e.*, efficiently traversing phylogenetic space, we expect the indicators to be flipping between 0 and 1 as frequently as possible, the maximum frequency depending on f_i. Let $\delta_i = \Delta(\boldsymbol{X}_i)$, where $\Delta(\cdot)$ a function that counts the number of state transitions in $\boldsymbol{X}_i$. Then $\sigma_i = \delta_i / J_i \in [0, 1]$ is a score that measures the relative efficiency of sampling by comparing how how many transitions happened compared to the theoretical maximum. This metric is quite similar to the Split Swap Rate of Höhna and Drummond (2008).

If the data are very informative, it is possible that some clade will have very high or very low posterior probabilities, meaning their indicators might never change. These clades are not interesting for assessing mixing and I therefore restrict attention to clades with $0.4 \leq f_i \leq 0.8$. This choice of "interesting clades" is tailored towards single chains. When multiple chains are available, Warren et al. (2017) propose tracking clades that have the highest changes in frequency across chains as these are more likely to be problematic. A multi-chain version of the switching score could be devised to complement this latter approach by measuring mixing in addition to diagnosing convergence issues.

It should be noted that under mild regularity conditions, the techniques described in Section 3.2.1 of Chapter 3 can be applied to binary variable such as the clade frequencies, hence the importance of the clade-switching approach is unclear. One could, for instance, compute the univariate ESS for each clade of interest as a way of quantifying mixing, taking the minimum across clades as a conservative metric, or the average as a more balanced statistic.

Both the clade switching score above and univariate ESS ignore the non-trivial dependence structure between clades and thus might not accurately reflect chain mixing. In principle, the multivariate ESS (mESS, Vats et al. (2015)) described in section 3.2.1 could also be computed as a global metric that takes correlation between clades into account. Understanding the correlation structure between clades

is therefore important and here I offer a sketch for a more complete characterisation of the correlation structure under the coalescent prior.

The correlation structure of any sample of clades $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$ is non-trivial, due to clades being incompatible or nested. Consider for example the clades $c_i = \{A, B, C\}|\{D, E\}$ and $c_j = \{A, B, C, D\}|\{E\}$: they are incompatible and hence would have 0 probability of co-occurrence, leading to a correlation $\rho_{ij} = -\frac{f_i f_j}{\sqrt{f_i(1-f_i)f_j(1-f_j)}}$. Another situation is when one clade is contained within the other , e.g. $c_i = \{A, B, C\}|\{D, E\}$ and $c_j = \{A, B\}|\{C, D, E\}$, hence $c_j \subset c_i$. In this situation, however it is not possible to write ρ_{ij} directly from the marginals f_i and f_j. We need to introduce the conditional frequencies $f_{01} = Pr(\mathbb{I}_i = 0|\mathbb{I}_j = 1)$, $f_{10} = Pr(\mathbb{I}_i = 1|\mathbb{I}_j = 0)$ and then we can derive:

$$\rho_{ij} = \frac{V(1-f_i) - f_i W}{\sqrt{(V+W)(1-(V+W))f_i(1-f_i)}}, \tag{3.11}$$

with $V = f_i - f_{10}(1-f_j)$ and $W = f_{01}f_j$.

Under the prior, we can compute the probability of a clade c_i by noticing it depends only on its size $|c_i|$ (Brown, 1994, Eq. 14):

$$f_i = \frac{2|c_i|!(n-|c_i|)!(n-|c_i|-1)!}{n!(n-1)!} \sum_{i=1}^{n-|c_i|} \frac{i(n-1-i)!}{(n-1-|c_i|)!} \tag{3.12}$$

For any two clades c_i and c_j such that $|c_i| + |c_j| < n$, one can calculate $f_{1,1} := Pr(\mathbb{I}_i = 1, \mathbb{I}_j = 1)$ using theorem 4.5 in Zhu et al. (2011) and then compute $f_{10} = (f_i - f_{1,1})/(1-f_j)$ using the law of total probability. Since $f_{01} = (1 - f_i - (1-f_{10})(1-f_j))/f_j$, it should be possible to compute all the correlation coefficients induced by an uniform prior on topologies, which is the case of the coalescent prior with contemporaneous tips[7].

[7]For serially-sampled tips, the number of trees is different as shown in Figure 1.3 (Gavryushkina et al., 2013) and I suspect the results from Brown (1994) would have to be modified. It is unclear to me whether this would be a trivial task.

By exploiting the combinatorial regularities of clade space (e.g. we know exactly how many entries in the clade correlation matrix have, say, $|c_i| = 4$, and $|c_j| = 2$) one can derive theoretical properties of the correlation matrix, such as its determinant. We know there will be $2^n - n - 2$ non-trivial entries in the matrix and it is possible to compute how many entries are negative (*i.e.* how many clades are incompatible). My main point here is that if one were to compute, say the mESS for a collection of taxa in hope that this would account for their correlation structure, it would be useful to understand what to expect *a priori*. I conjecture it may be possible to show that the correlation matrix induced by the prior is ill-behaved and does not allow for computing mESS – initial investigations using the `multiESS()` from the **mcmcse** package (Flegal et al., 2017) in R resulted in singular correlation matrices. Figure 3.1 shows the prior correlation matrix for $n = 5$ with contemporaneous tips. A comparison between clade switching, univariate ESSs and mESS for quantifying mixing in clade space could be an interesting future project.

3.2.5 Multi-dimensional scaling

One way of visualising phylogenetic space is by employing lower-dimension projections that attempt to embed it in Euclidean space. A very popular technique for this purpose is multi-dimensional scaling (MDS, Hillis et al. (2005)) of phylogenetic distances, in which the idea is to construct a new space with a few (usually two) dimensions that captures most of the features (variability) of the original space. Suppose $\boldsymbol{D}$ is a phylogenetic distance matrix where D_{ij} is the distance – under a particular metric d_σ – between two phylogenies i and j. MDS proceeds by finding points $\boldsymbol{x} = \{x_1, x_2, \ldots, x_P\}$ such that a **stress function** (Kruskal, 1964)[8]:

$$S_D(\boldsymbol{x}) = \sqrt{\sum_{i \neq j = 1}^{P} \left(D_{ij} - |x_i - x_j|\right)^2}, \tag{3.13}$$

[8]Other stress functions are possible, I keep the Kruskal-1 function here for simplicity and comparability with Hillis et al. (2005).

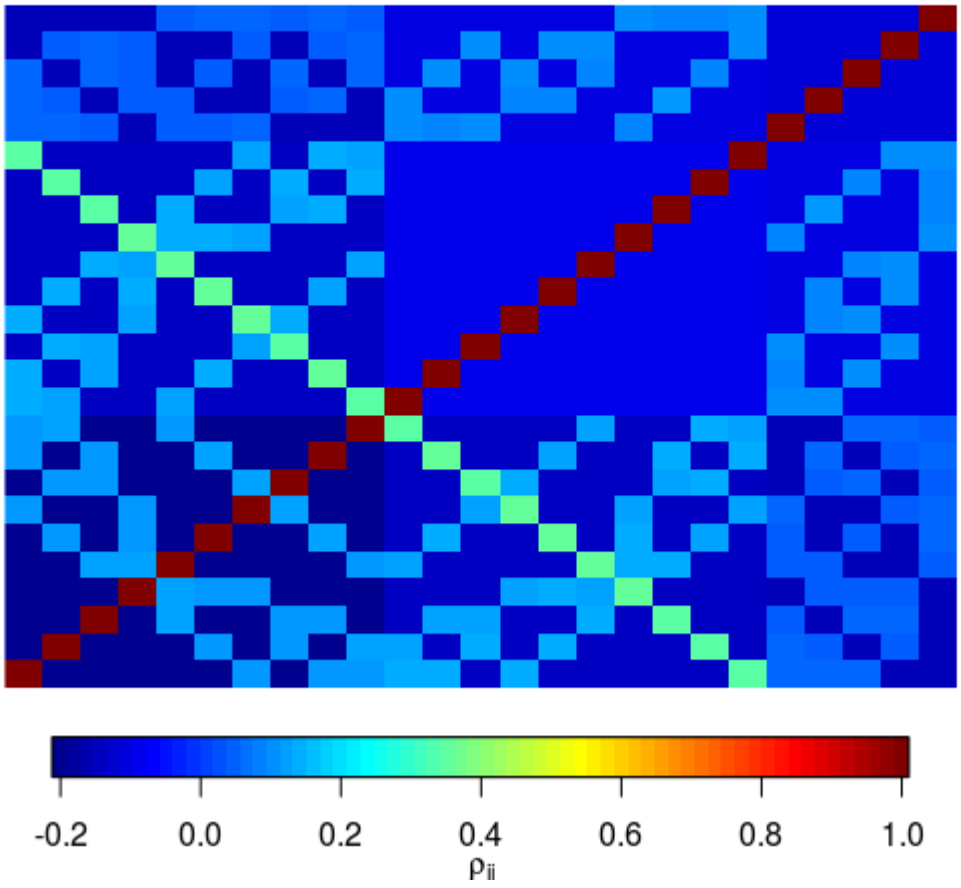


Figure 3.1: Clade correlation matrix (coalescent prior), $n = 5$. I sampled 10, 000 trees under the constant population size coalescent prior with contemporaneous tips using BEAST and computed the correlation matrix between clades of sizes between 2 and 4. To facilitate visualisation of the structure in the matrix, I sorted clades by size (grows from bottom to top and left to right).

is minimised. As pointed out by Hillis et al. (2005), under certain conditions this ensures the new space does not contain large distortions from the distance matrix and hence the new space can be used as representation of phylogenetic space. Once

we have an optimal solution computed from the observed distances, $\boldsymbol{x}$, we can use it to produce visualisations.

While the original paper by Hillis et al. (2005) employed Robinson-Foulds to compose $\boldsymbol{D}$, Jombart et al. (2017) generalise that approach by developing an R package to aid MDS visualisation under different metrics, with special focus on the KC metric (Kendall and Colijn, 2016). In addition, Jombart et al. (2017) also provide ways of extracting representative trees from a large sample, which reflect median points around which trees cluster. This might prove particularly useful when finding and characterising *modes* in phylogenetic space.

In order to enable meaningful comparisons between runs, I employ the Procrustes method to obtain an optimal rotation/scaling of the resulting distance matrices such that they are as compatible as possible. This helps minimise distortions caused by stochastic variation and allows MDS projections from different sets of phylogenies to be overlaid for comparison. I used the function `procrustes()` from the R package **vegan** (Oksanen et al., 2018) to obtain Procrustes-transformed matrices.

Limitations

As noted in Section 3 of Willis and Bell (2017), phylogenetic space cannot be completely embedded in Euclidean space, hence any procedure that produces a mapping $\boldsymbol{\Phi} \rightarrow \mathbb{R}^P$ will invariably lead to the loss of some information. A comprehensive assessment of the relative merits of different stress functions and phylogenetic metrics is lacking, and unfortunately outside the scope of this chapter. Moreover, since 2-d visualisations are so much easier to understand, practitioners might be tempted to only consider the first two coordinates of the transformed space, what might in turn lead to missing important features in the data.

3.2.6 Graph (network) analysis of tree space

Another useful way of representing the discrete (tree) component of phylogenetic space is equipping it with a metric d_σ and constructing a graph $G_\delta(V_n, E_n)$ where each vertex corresponds to a tree topology and there is an edge between two edges (trees) if they are a distance $d_\sigma \leq \delta$ apart under the chosen metric. Thus, for this "meta-graph", $|V_n| = |\mathbb{F}| = (2n-3)!!$ and $|E_n| \leq |V_n|(|V_n|-1)/2$, although this last bound is very crude and can be refined for a given d and δ. Common choices for metric include the nearest-neighbour interchange (NNI) and subtree prune-and-regraft (SPR) distances. In particular, Whidden and Matsen (2015) show how one can construct the SPR graph of a sample of trees and then use graph-theoretic tools to quantify exploration of phylogenetic space. Starting from a sample of trees $\boldsymbol{\tau} = \{\tau_1, \tau_2, \ldots, \tau_M\}$, the analysis pipeline proposed by Whidden and Matsen (2015) can be summarised as follows:

0. Rank the elements in $\boldsymbol{\tau}$ by their posterior probability (frequency), creating a set $\boldsymbol{\tau}_{\text{top}}$, also determine the most probable tree, $\tau_{\max}$;
1. Keep only the $m = \min(4096, |\boldsymbol{\tau}_{\text{top}}|)$ first elements;
2. Compute the $m \times m$ matrix of SPR distances between the samples, $\boldsymbol{D}$, and construct a graph $\boldsymbol{G}$ where each tree is a vertex and two vertices i and j are connected by an edge if $D_{ij} = 1$, that is, if the two phylogenies are one SPR operation apart from each other;
3. Identify clusters of high-probability trees by iteratively clustering trees until $K = 8$ clusters are obtained;
4. Visualise the resulting graph annotating the vertices with posterior probabilities and distance to the most probable (or mode) tree.

Whidden and Matsen (2015) also propose three metrics – which can all be computed with a single pass on $\boldsymbol{\tau}_{\text{top}}$ – to quantify exploration of tree space:

1. Mean access time (MAT): average number of iterations required to visit any two trees;

2. Mean commute time (MCT): average number of iterations required to go from $\boldsymbol{\tau}_{\text{top}}$ to each of the other high probability trees and back;

3. Round-trip cover time: starting from $\tau_{\max}$, the average number of iterations necessary to visit each high probability tree and then return to the mode tree.

Limitations

One of the advantages of building $\boldsymbol{G}$ using SPR distances is that many of its theoretical properties are better studied (Whidden and Matsen, 2017). On the other hand, focussing on SPR distances disregards branch lengths, which is not always desirable, specially when dealing with time-calibrated phylogenies (see below). Perhaps a more serious flaw with the SPR-graph approach is that for large n, each tree visited in the MCMC is likely to be unique, and hence one cannot construct $\boldsymbol{\tau}_{\text{top}}$ based on posterior frequencies.

3.2.7 Effective sample size and potential scale reduction factor for phylogenies

Recently, researchers have also extended the concepts of effective sample size (ESS) and potential scale reduction factor (PSRF) to tree topologies, taking advantage of the fact that these quantities are defined w.r.t. the L^2 in Euclidean space. Lanfear et al. (2016) propose two ESS-like metrics: (i) the **pseudo-ESS**, n_{eff}^P and (ii) the topological **approximate ESS**, n_{eff}^T. If we define τ_f to be a focal tree and let $\boldsymbol{L} = \{l_1, l_2, \ldots, l_M\}$, $l_i = d_\sigma(\tau_i, \tau_f)$, the pseudo-ESS is simply the ESS of $\boldsymbol{L}$ computed as defined above (Eq 3.3). In other words, this measure attempts to reduce phylogenies to a univariate continuous quantity for which the methods discussed in Section 3.2.1 can be applied. An advantage of this approach is that

it admits any choice of metric d_σ, which in turn means one can use metrics that account for branch lengths and topology simultaneously. A perhaps more principled approach is to try to directly estimate the autocorrelation spectrum for phylogenies. Let d be the squared distance between two independent phylogenies. The expected value of d with N independent samples is $(N(N-1)/4N^2)d$ and we can estimate N (as n_{eff}^T) when our sample has M sequential observations by noting that

$$\frac{N(N-1)}{4N^2} = \frac{\sum_{i=1}^{M-1}\sum_{k=1}^{\min(m,M-i)} f(k) + \frac{1}{2}(M-m+1)(M-m)d}{2M^2}, \tag{3.14}$$

$$N = \left[\frac{2\sum_{i=1}^{M-1}\sum_{k=1}^{\min(m,M-i)} f(k) + (M-m+1)(M-m)d - M^2}{M^2}\right]^{-1}, \tag{3.15}$$

where $f(k)$ is the squared distance between two samples at sampling interval (lag) k and m is the minimum sampling interval at which any two samples are independent, *i.e.*, the point where the autocorrelation function reaches an asymptote. Lanfear et al. (2016) suggest using an exponential model to estimate the autocorrelations, $f_a(k) = d(1-\exp(k/a))$, where a is a shape parameter. Implementations of both these functions are available in the **RWTY** package (Warren et al., 2017).

The idea of treating squared distances as a surrogate for variance as employed above can be further extended to compute a PSRF-like for multiple samples of phylogenies. Whidden and Matsen (2015) propose a "topological Gelman–Rubin-like convergence diagnostic", $\hat{R}'$, that is very similar to (3.7) with B and W replaced by:

$$W' = (M(M-1))^{-1}\sum_{k=1}^{K}\sum_{j_1=1}^{M}\sum_{j_2=1}^{M} d_\sigma(\tau_{kj_1},\tau_{kj_2})^2, \tag{3.16}$$

$$B' = \left((K-1)KM^2\right)^{-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\sum_{j_1=1}^{M}\sum_{j_2=1}^{M} d_\sigma(\tau_{i_1j_1},\tau_{i_2j_2})^2, \text{ hence} \tag{3.17}$$

$$\hat{R}' = \sqrt{\frac{(M-1)W' + B'}{MW'}}. \tag{3.18}$$

As with the original PSRF, $\hat{R}'$ approaches 1 as the K independent runs converge, and $\hat{R}' < 1.1$ cut-off for convergence could be adopted.

Limitations

According to Lanfear et al. (2016), the pseudo-ESS can be sensitive to the choice of focal tree. While that could be remedied in principle by choosing, for instance, the maximum clade credibility (MCC) phylogeny as τ_f, it is unclear to me what biases this could introduce. The original formulation of (3.18) by Whidden and Matsen (2015) considered only SPR distances, but this can be easily generalised to other distances that include branch lengths. A minor point about these diagnostics is that due to their recent development, in-depth theoretical and empirical studies are still lacking. For instance, the choice of metric might play an important role in the power to detect convergence problems, but this aspect remains to be investigated.

3.2.8 A word of caution

As Charlie Geyer points out in the very quaintly named web page "On the Bogosity of MCMC Diagnostics"[9], all known diagnostic measures have serious shortcomings and, in his words, can detect only "... obvious, gross, embarrassing problems that jump out of simple plots". While I do not completely subscribe to Geyer's view, I agree that subtle problems such as small biases in the Hastings ratio (see e.g. Holder et al. (2005)) or funnel-like effects of posterior (prior) correlation are likely to remain undetected. Therefore a word of caution is warranted: even when one fails to detect convergence problems (Cowles et al., 1999), they may very well still be present in the form of subtle biases, the impact of which on inferences drawn from the MCMC samples is hard to predict.

[9] Available from `http://users.stat.umn.edu/~geyer/mcmc/diag.html`, accessed on 2018-02-18.

3.3 Accommodating time-calibrated phylogenies

In this Chapter I provide a first attempt at an unified workflow for Bayesian estimation of time-calibrated phylogenies (TCPs) with special focus on phylodynamic inference. As discussed previously, TCPs are special objects in that the branch lengths are measured in units of calendar time. In what follows I detail my investigations into several issues related to accommodating time-calibrated phylogenies and diagnosing the convergence of MCMC runs from BEAST. Considering the methods currently available, the main difficulties faced when exploring the space of TCPs are:

- Size: the phylogenies used in phylodynamic studies have n of the order of hundreds to a few thousands (see below);
- Branch lengths: in time-calibrated phylogenies branch lengths are of crucial importance and hence cannot be ignored;
- Sampling structure: TCPs usually have serially-sampled tips/leaves. The distribution and range of the sampling times imposes constraints on the phylogenetic space, leading to "rugged" posterior distributions (Brown and Thomson, 2018) – see also Figure 3. in Möller et al. (2018) and discussion therein.

Most routines in AWTY are designed for unrooted trees and implicitly assume a relatively small number of taxa ($n < 50$). Moreover, the MDS routines presented in Hillis et al. (2005) and available in the **RWTY** package use the Robinson-Foulds distance which does not capture branch length differences and hence is not appropriate for time-calibrated phylogenies if used in isolation. The routines available in the **RWTY** package also use RF as the default metric, and as far I am aware, the only metric that includes branch lengths available in the package is path distance (BS, see Section 3.1.1). Here I relax this by including many other

metrics (see below), for which the approximate ESS of Lanfear et al. (2016) can be computed using equation (3.14). As the number of taxa grows, it gets progressively harder to track clades, both from a statistical and a computational point of view. Computationally, it becomes cumbersome to compute indicators for all clades in a run, in addition to plotting clade frequencies. This latter problem can be tackled by only paying attention to clades with a particular (posterior) frequency. Statistically, however, the space of clades presents some non-trivial correlation structure that might make it hard to obtain reliable global indicators of convergence and mixing. In Section 3.2.4, I provide a more detailed discussion of these issues. Nonetheless, in the interest of consistence and comparability with previous approaches I include diagnostics of convergence and mixing in clade space.

In a phylodynamic analysis context one may be chiefly interested in a set $\boldsymbol{\theta}^{\star}$ parameters, which might include quantities such as the reproductive ratio R_0 (Stadler et al., 2011) or the wave front velocity of an epidemic (Lemey et al., 2010; Pybus et al., 2012). It is therefore important to study the behaviour of the chains for $\boldsymbol{\theta}^{\star}$, as well as account for the correlations between parameters. Fortunately, there are a plethora of tools designed to diagnose convergence for continuous parameters, many of which are available in the R package **coda** (Plummer et al., 2006) and the GUI application Tracer (Rambaut et al., 2018).

The pipeline/workflow proposed here can be summarised as follows:

1. Run (at least) three independent chains for M iterations each, keeping M_t phylogenies;
2. Compute univariate ESS, PSRF and mESS for $\boldsymbol{\theta}^{\star}$;
3. (Sub)sample a number $K < M_t$ of trees and compute $(K \times K)$ distance matrices under different metrics;
4. Perform MDS on the distance matrices and use them to visualise phylogenetic space (see Figure 3.2b);

5. Using the same distance matrices, compute the approximate ESS (Lanfear et al., 2016) for each run under different metrics;

6. Compute clade frequencies and indicator matrices and calculate clade ESS, clade switching and average standard deviation in split/clade frequencies (ASDSF);

Specialised computer programs

Most modern phylodynamic studies include data sets with hundreds to a few thousands of samples (taxa) and the resulting phylogenies strain the capacity of most existing packages (including AWTY and RWTY). Computationally, one of the main bottlenecks is loading trees into memory – which is quite slow in R, for instance. I instead do the tree-processing externally using specialised classes in BEAST (`dr.app.tools.TopologyTracer`)[10], controlled using simple bash scripts. I then wrote custom R functions to transform the output of `TopologyTracer` so that it could be analysed using heavily modified functions in the package **RWTY** Warren et al. (2017). The same strategy can be adopted when processing clade frequencies: I used the class `dr.app.tools.TreeSummary` to compute clade frequencies and construct the indicator matrix and then modified functions from the **RWTY** package for plotting and presenting the results. As mentioned above, I employ the packages **mcmse** (Flegal et al., 2017) and **coda** (Plummer et al., 2006) to compute (univariate and multivariate) effective sample sizes and potential scale reduction factors for continuous parameters. This is done to keep all computations contained in the R environment, which allows the whole workflow to be encapsulated into a (R)Markdown document which can be rendered to html and/or PDF (see Figure 3.2). A suite of R and bash scripts – as well as the RMarkdown report – to perform these tasks is available from `https://github.com/maxbiostat/BEAST_`

[10]Implemented by Andrew Rambaut and Guy Baele (Leuven) with input and testing from me.

`convergence_pipeline`[11]. For convenience, the analysis of the "poor" runs (see below) is included in Appendix A as an example.

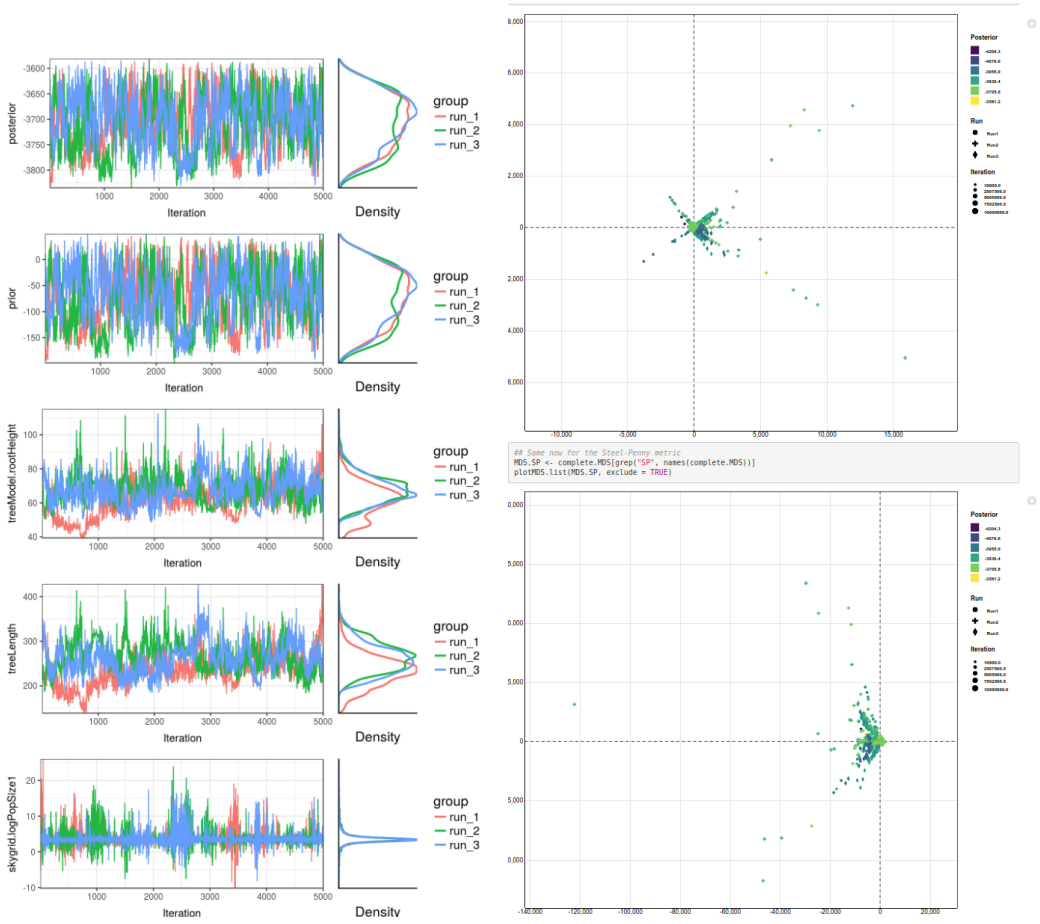
3.4 Data sets

Here I will analyse the same data sets described in Table 2.1 in Chapter 2. In particular, `Dengue4`, due to its manageable size but relatively high complexity (multiple partitions, temporal structure, etc), will be used to exemplify the full breadth of convergence diagnostics available and how they may be combined for maximum efficacy. I will analyse MCMC runs resulting from empirical analyses of these data sets to investigate the characteristics of phylogenetic space for serially-sampled data modelled under phylodynamic models. For the results in Section 3.5.1, I constructed a deliberately bad MCMC sampler to obtain a poorly mixing chain. To achieve this I employed only the `WideExchange` operator – with small weight – to sample tree topologies. Since this operator is not "tunable" and has a very low acceptance probability (0.021 for this data set), the chain mixes poorly. These poorly mixing (Poor) runs were then compared with runs using the default settings in BEAUti, the GUI configuration file maker for BEAST. In addition, I ran the pipeline with `STL` replacing all tree-related operators with the `SubTreeLeap` operator described in Chapter 2. All runs had 10 million iterations and a thinning interval of a thousand states.

3.4.1 Simulated data

I also created a simulated data set in order to have a baseline where we know the ground truth. I first obtained a 50 taxa subtree with uniformly temporally sampled tips from a larger EBOV phylogeny (Dudas et al., 2017), henceforth called "empirical tree". To assess the effect of tree shape, I simulated a coalescent tree with the same

[11] If these functions prove to be sufficiently useful to other researchers, I may build an R package.



(a) **Continuous parameters** (b) **MDS**

Figure 3.2: Screen capture of the proposed MCMC diagnostics pipeline. One can check trace plots – as well as investigate univariate and multivariate ESS – for continuous parameters (a) and study exploration of phylogenetic space by visualising MDS (b) – note these latter plots are made interactive using the spreadD3() function in R. All routines have been written in R and organised into an RMarkdown document that generates an html report that can easily explored by the user to diagnose problems with her MCMC runs. See Appendix A for more.

tip-date sampling structure under a constant population model with $N_e = 10$, which I will call "coalescent tree". I then simulated alignments of 1,000, 5,000 and 10,000 sites down both trees with a substitution rate of 10^{-3} substitutions per site per

year, generating 6 alignments. For each alignment, I then ran three independent runs (replicates) of both STL and the default/classic mix. To explore convergence times and obtain more stable estimates of the posterior, each combination of data set and transition kernels was run 10 million and 100 million iterations.

3.5 Analysis

3.5.1 Combining diagnostic measures

The first set of results I will present is the development of an analysis "pipeline" for MCMC diagnostics in Bayesian phylogenetics. In particular, I will use the Dengue serotype 4 *env* (`Dengue4`) – 17 taxa, 1485 sites – to showcase how many convergence diagnostic measures can be employed in conjunction as part of the Bayesian phylogenetic workflow. Appendix A shows the analysis of the poorly mixing runs using the steps and computer programs described above. In this section I focus on comparing some results for three MCMC schemes: poorly mixing, default settings and `STL`.

First, I present the results of convergence diagnostics for continuous parameters, usually obtained with the graphical tool Tracer in the context of BEAST analyses. Here however I compute some additional quantities not available in Tracer, such as multivariate ESS and potential scale reduction factor (PSRF). Results in Table 3.1 show that the poor runs have low univariate ESSs for continuous parameters highly dependent on the tree (e.g., mean evolutionary rate, `meanRate`), whereas ESS are in the low thousands for parameters such as the transition transversion rate (κ). For this particular experiment and with no optimisation, `STL` shows comparable if slightly inferior performance when comparing raw univariate ESSs. The PSRFs show that for many of these parameters the three independent runs do not converge to the same point (PSRF $>$ 1.1), be it individually for each parameter or globally across parameters (multivariate PSRF $>$ 1.1). While the `STL` runs seem more

consistent, as indicated by the slightly lower PSRFs, with only three runs (chains) per MCMC scheme this difference cannot be reliably established (see Chapter 2 for more thorough comparisons). The multivariate ESSs might at first glance give the impression of good performance, but in fact all of them fall well short of the minimum ESS required for reliable inference (see caption in Table 3.1). Notice that while this is specially the case for the poor runs (as expected), none of the runs across MCMC schemes achieve the minimum ESS (8831) after 10 million iterations (see Discussion).

Considering mixing and convergence in phylogenetic space through the use of specially tailored diagnostics presented in Table 3.2, shows that while the approximate ESS (Lanfear et al., 2016) does capture major differences in mixing, it is inconsistent, both within and between metrics. As an example, take the tree ESS with the Steel-Penny (SP) metric for the Poor runs: for one of the chains, it achieves its maximum value of 1001, when it clearly cannot be the case that the sample comprised 1001 independent samples considering all available evidence. The clade switching score (CSS) and mean ESS for clade indicators seem to discriminate well between the Poor runs and the default (and `STL`) ones. ASDSF is also above the common threshold of 0.10 used for convergence (Ronquist et al., 2012); `STL` seems to lead to slightly better performance according to this metric.

I now move on to explore some more specific questions about phylogenetic space and its representations.

3.5.2 Representation of phylogenetic space under different metrics

While multi-dimensional scaling can be useful for visualising phylogenetic space, the issue of which metric to employ remains. Since each of the many available metrics captures distinct features, one has to analyse MDS projections under different metrics in order to have a better grasp of the geometry of phylogenetic space. In

Figure 3.3 I show the MDS projection of Robinson-Foulds (RF) distances between MCMC samples for the large Ebola virus data set (`EBOVa`) – the analysis detailed in Chapter 2, see Figure 2.10. It is clear that run 3 for the `STL` operator (see Chapter 2) is distinct from the others. In addition, one can see that the last samples of run 2 `STL` (larger green points) are closer to the region visited by the default runs and `STL` run 1. These results show that the projection using the Robinson-Foulds capture important features of phylogenetic space, leading to clearly separated clusters of trees with different likelihoods.

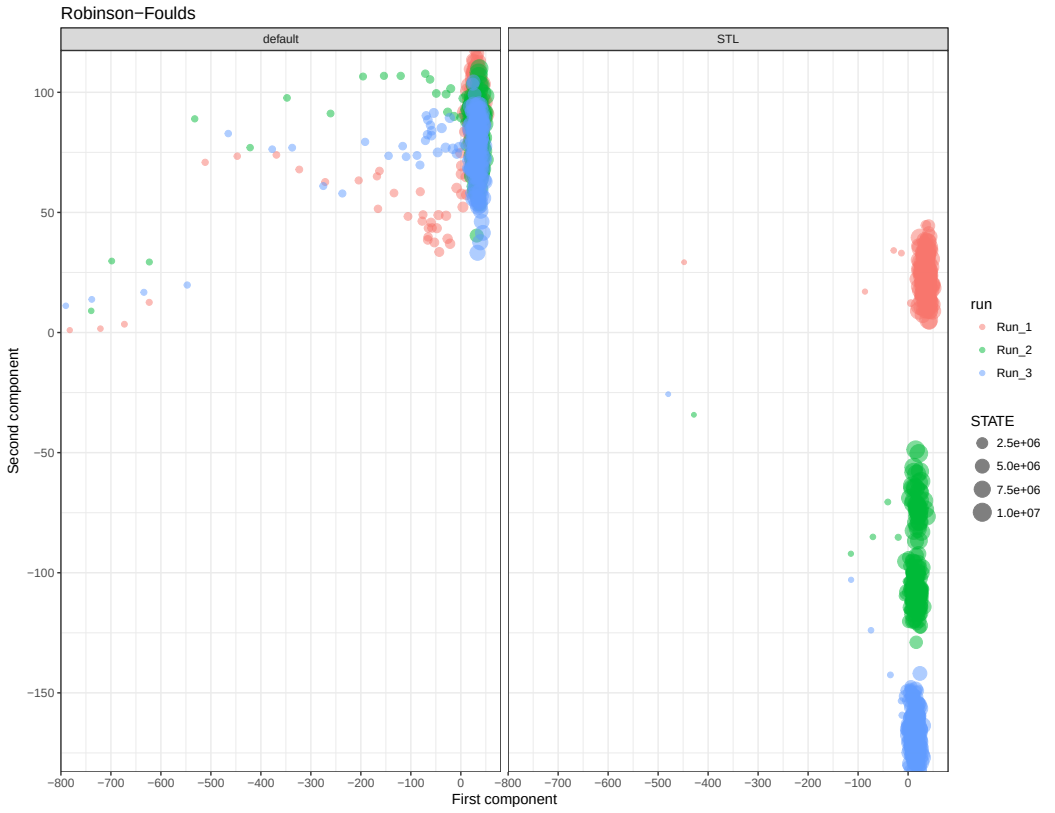


Figure 3.3: MDS projections for the full EBOV 1610 taxa data set (Robinson-Foulds distances). I sampled 200 equally spaced trees from each chain (1200 trees in total) and computed the Robinson-Foulds distance (Robinson and Foulds, 1981) for every pair of trees. Colours relate to the seed used in the pseudo random number generator, ensuring each run starts from the same point. Please see Figure 2.10 in Chapter 2 for more details. The size (radius) of the dots is proportional to the state of the chain.

Importantly, however, these differences are not captured by other metrics. The projections under the Kendall-Colijn (KC) metric in Figure 3.4 shows no obvious clustering of the runs, even though we know different runs explore trees with very different likelihoods. In panel (a), I show the MDS projection of KC metric with $\lambda = 0$, that considers only topological differences between phylogenies, but the projections do not show any obvious clustering of the runs, unlike the RF metric (Figure 3.3). The projection with $\lambda = 1/2$ in panel (b) displays the same pattern. This seems to suggest that MDS under this metric does not allow one to visualise the different modes in phylogenetic space, or, in other words, that KC is too smooth to capture the differences, at least for this data set.

In Figure 3.5 we can again notice the lack of differentiation between the runs, suggesting the SP metric shares the "smoothness" displayed by the KC metric. Taken together, these results suggest that great care needs to be taken when visualising phylogenetic space because important features might not be easy to distinguish under many (most) metrics.

One question seldom approached in the literature is the quality of MDS projections of phylogenetic space. As mentioned above, it is standard practice to keep the first two components for plotting and analysis, but if this 2-d projection is to be a faithful representation of the space of interest we need to ascertain that the first two components do indeed capture most of the variation. A way of analysing the quality of the projections is to plot the scaled eigenvalues against the component number. Ideally, the first few components capture most of the variation, resulting in an "elbow-shaped" plot. In contrast, a flat plot suggests that variation is spread across many components and hence restricting attention to the first two can be misleading. In Figure 3.6 I show these plots for a range of real-world data sets (see Table 2.1 in Chapter 2). A consistent result across data sets is the plot for the Robinson-Foulds (RF) distance, which is mostly flat and suggests that for this metric it is not sufficient to look at the first few components in order to get a good picture of the variation in the sample.

3.5.3 Typical set for phylogenies

One question one might ask is, what does phylogenetic space look like when we know the right answer? In this section I offer a first stab at this question by analysing a simulated data example with 50 taxa (Section 3.4.1). I present trace plots of the distance to the true tree and MDS projections of the posterior distributions under various metrics, marking the maximum clade credibility (MCC) trees obtained from each run of each operator mix and the true tree used to generate the data. In Figure 3.7, I show the MDS projection of the RF distances between trees sampled using three MCMC schemes ("operator mixes" or MCMC schemes, see Chapter 2), under two generating models: a tree drawn from the coalescent and an empirical tree extracted from a real-world data sets. We can see that as the amount of data increases, the posteriors become more concentrated around the true value, as do the MCC trees, as expected. While this behaviour is present for both true trees (coalescent or empirical), the posteriors for the coalescent generating tree seem to be more variable.

Under the KC metric with $\lambda = 0$ (Figure 3.8), we observe a very similar pattern, but for this data set this representation more clearly shows the topological modes and how the MCC trees approach the correct mode (that contains the true tree) as the amount of data grows (rightmost bottom panel) . Also from this panel we see that the posterior for the empirical generating tree also show markedly separate modes when compared to coalescent generating tree. Since this representation captures only topological features, I claim that the differences must be due to the implicit interaction between branch length distribution and topology. An interesting observation is that the default operators sometimes lead to samples outside the typical set even when the chain is supposed to have converged to the target (rightmost top panel).

In keeping with the results in Section 3.5.2, the Steel-Penny metric leads to MDS projections that appear smoother (Figure 3.9). Also noteworthy is the amount of

samples outside the typical set observed for the default kernels (small purple dots). Taken together with Figures 3.7 and 3.8, these results show progressive concentration of the posterior around the true value with alignment size (number of sites), even if posteriors for the empirical generating tree seem less smooth and harder to sample from.

When we know the true tree, we can also study phylogenetic space and its exploration *via* MCMC by computing the distance to the true tree and tracking how the distance changes through the chains as well as the resulting distributions. In this simulated example we can exploit the fact that we know the true tree to the visualisation/analysis problem to essentially one dimension. We can look at the resulting distributions as if they were univariate targets, which in turn facilitates visualisation and intuition-building, in addition to making it possible to apply a plethora of statistical methods developed for the analysis of univariate distributions. While in practice we do not know the true tree, these simulations are useful for understanding the behaviour of transition kernels and MCMC in general.

In Figure 3.10 I show the results of this analysis for the 50 taxa simulated example discussed above, for topological metrics (RF and KC) in order to show the combinatorial (discrete) multimodality of phylogenetic space. The first pattern to notice is that the number and location of peaks changes as the amount of information increases (colours). Secondly, I highlight how running the chains for longer leads to better defined peaks, even though for this simple example[12] 10 million iterations seem to be adequate to find and explore all of the detected modes. Also noticeable is how diffuse the target for the coalescent tree is for a small alignment size (one thousand sites) under the KC metric (Figure 3.10, panel D). While the target intermediate alignment shows bimodality, the target for 10 thousand sites shows three modes, one much bigger than the other two. Overall the empirical generating

[12]Recall that in this example the parameters are inferred under the generating model,*i.e.* there is no model misspecification.

tree leads to more complex posteriors, specially for the intermediate alignment size (panel B, red curve).

When we turn attention to metrics that incorporate branch length information (Figure 3.11), another interesting pattern emerges: while the SP metric leads to smooth targets, the KC metric shows distinct multimodality for the empirical generating tree. It is difficult to say whether these observed differences are due to some underlying fundamental distinction or just an artefact specific to the two trees used – there were no replicates at generating tree level of the experiment. This result calls for further investigation into the inherent differences between empirical trees, *i.e.*, trees that are estimated from data encountered in practice, and their coalescent counterparts (see Section 3.6).

Figures 3.10 and 3.11 also illustrate how casting the phylogeny sampling problem as a univariate sampling problem allows one to more clearly compare MCMC strategies. Under both representations it is clear that `SubTreeLeap` leads to virtually identical samples to those obtained with the default set of transition kernels, increasing confidence in the correctness of its implementation (see Chapter 2, Section 2.2.3). Moreover, it also seems to find and sample from all the detected modes, even for complex distributions such as those in panel B of Figure 3.10.

While reducing the problem to a univariate quantity is an attractive idea (see, e.g. "Method 1" in Lanfear et al. (2016)), it is generally not possible in practice since we do not know the true tree and the choice of focal tree is arbitrary. In the context of the development of phylogenetic transition kernels, however, I believe this method can be useful in constructing univariate representations of the target distribution, and thus allow us to leverage a vast body of theory available for assessing performance and correctness. See Section 2.5.3 in Chapter 2 for more details on how this framework can be applied to evaluate MCMC mixing.

3.6 Final remarks

In this chapter I have sought to supplement the toolbox of diagnostic measures for MCMC in phylogenetic space with a few new tools and other tools seldom used in the field. I by no means claim to have provided the first workflow of this kind, however. Previous studies such as Hillis et al. (2005), Lakner et al. (2008), Nylander et al. (2008) and Warren et al. (2017) have touched on many of the aspects tackled here. The research reported here is an attempt at combining previous approaches with new metrics while overcoming some of the technical hurdles imposed by large time-calibrated phylogenies impose.

The results for the "poor" runs (Table 3.1) are not surprising since the chain was deliberately designed to mix poorly in phylogenetic space. A perhaps more subtle point to be made, however, is that since parameters vary regarding their inherent dependence on the phylogeny, computing global convergence metrics including all parameters might mask convergence problems. This could be easily fixed by separating parameters into blocks when assessing convergence; more phylogeny-dependent parameters could be paid special attention. Of course it is not always easy to know how dependent on the underlying phylogeny a parameter is; further research is needed in order to determine whether/how to do parameter blocking. The multivariate ESS results show that when accounting for correlations between parameters none of the tested MCMC schemes achieves a sufficient number of samples. This is intimately linked with the common recommendation of declaring a run acceptable if it achieves marginal univariate ESSs larger than 200 for all parameters. The practical relevance of this rule is yet another topic for future research.

Regarding diagnostics specifically designed for phylogenies, the results in Table 3.2 suggest a lack of agreement between approximate ESSs computed using different tree metrics. For instance, for the poorly mixing runs some ESSs computed using the clade difference (CD) and branch score (BS) are estimated as 1001, the maximal

value, suggesting these metrics might not be reliable for discriminating between poorly mixing runs and adequately mixing ones. These results are in tune with those presented in Section 3.5.2, which show that MDS projections under some metrics fail to highlight differences between runs and thus indicate severe convergence problems. In contrast, Figure 2.14 in Chapter 2 shows that pseudo-ESS seems consistent between metrics – the ranking between MCMC schemes is largely consistent –, at least in a setting where the focal tree is a summary tree from the tail end of three independent (long) golden runs. Clade-based metrics seem to provide more discriminating tools, albeit the experimental setup does not allow definitive claims to be made. As discussed in Section 3.2.4, correlation structure between clades complicates diagnostics based on marginal quantities, what seems to be corroborated by the fact that for the poorly mixing runs the mean univariate ESS for clade indicators was around 800, which could mislead researchers into inferring acceptable mixing.

These inconsistencies seem to stem from the non-standard nature of phylogenetic space, which admits many representations without any one way of depicting the space being canonical. The Billera-Holmes-Vogtman (BHV, Billera et al. (2001)) cubical complex representation is a good candidate with good statistical properties, but distances are hard to compute, while others such as Kendall-Colijn metric are easier to compute but suffer from poor statistical and/or theoretical properties. The KC metric for example suffers from an inherent scaling problem: when branch lengths and node indices are not commensurate, it is hard to calibrate λ in order to obtain interpretable results. A study in the same vein as Kuhner and Yamato (2014) to compare the relative efficiency of tree metrics specifically for time-calibrated phylogenies is sorely needed.

The attentive reader will have noticed that while I mention and describe the BHV (geodesic) metric and associated space, none of my results include this metric. This deserves explanation. While conceptually I believe the BHV representation of phylogenetic space to be the most complete, enjoying many desirable properties

(see Billera et al. (2001), St. John (2017) and Dinh et al. (2016)), despite recent advances in the computation of the BHV metric (Owen and Provan, 2011) it still remains quite hard to apply to large time-trees. Further programming work is needed to integrate the libraries provided by (Owen and Provan, 2011) into the framework described in this chapter. In addition, if we are to approach phylogenetic inference as a statistical problem (Holland, 2013), we need a representation that is statistically motivated so that results such as central limit theorems can be established. Under such a representation, concepts such as effective sample sizes and typical sets are easier to interpret and analyse. See Chapter 6 for a discussion of how BHV-based representations can be extended in that direction.

In summary, this chapter provides the following contributions/findings:

- New analytical tools for a robust Bayesian phylogenetic analysis pipeline;
- For a given data set, some metrics can completely fail to indicate important differences between runs (compare Figures 3.3, 3.4 and 3.5);
- Depending on the data set and metric, using just the first two components can be misleading (Figure 3.6);
- Approximate tree ESSs based on different metrics (distances) gave inconsistent results;

Several questions remain open however, which I intend to tackle in future research:

- How exactly does sampling affect (the exploration of) phylogenetic space? The results in Section 3.5.3 suggest that, for a highly idealised situation (relatively few taxa, no model misspecification) the typical set of topologies under most metrics is smooth and well-behaved, with posterior concentration around the true tree as alignment size increases and the MCC tree closer to the true data generating tree. It remains to be seen how model misspecification, for instance, changes these findings.

- Are there good universal – data set independent – cut-offs for ASDSF? What about the clade switching score? Tracking clades can be an efficient way of counteracting the superexponential growth in the dimensionality of the parameter space, since the space of clades grows much slower. However, non-trivial correlations between clades mean that naive metrics that do not take this correlation into account will fail to spot nonconvergence. The "holy grail" of MCMC for phylogenies convergence assessment is to find a cheap, preferably univariate measure that captures convergence with high sensitivity.

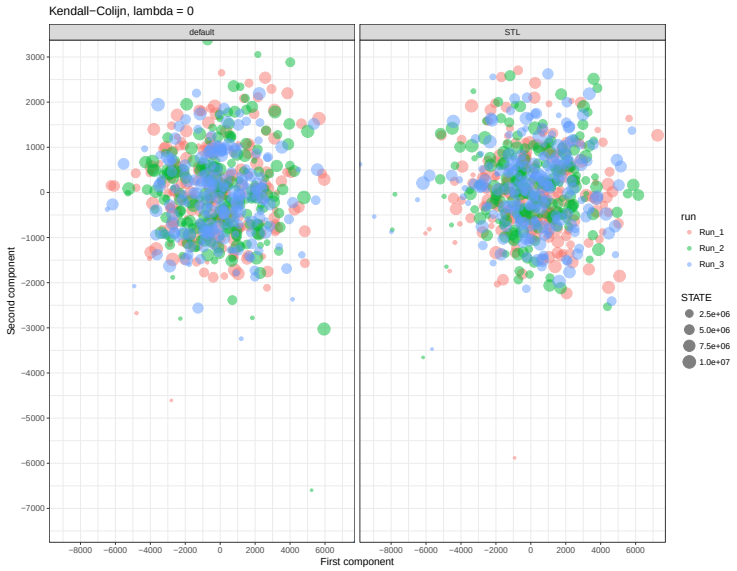
- Is it possible to elect one particular metric as more appropriate for the analysis of time-calibrated phylogenies? As mentioned above, a study extending the results of Kuhner and Yamato (2014) to time-calibrated phylogenies would be a good contribution. Such a study would have to consider how to incorporate the serially sampling structure of tips into the experimental design, seeing as the sampling dates of the tips impose constraints not only on topology but also on branch lengths (Möller et al., 2018).

Table 3.1: Convergence diagnostics for continuous parameters. I show the summary convergence diagnostics from the pipeline described in Section 3.3 for three MCMC schemes, running three chains for each. [1] - First (log) population from Skygrid. [2] - The ratio between the average rate and the standard deviation of rates across all branches. [3] - Covariance between rate assignments in the tree. [4] - Multivariate ESS as in Vats et al. (2015) and multivariate potential scale reduction factor (PSRF). The minimum multivariate ESS for all parameters considered should be 8831 according to the formula in (3.9).

	Poor				Default				STL			
Parameter	ESS 1	ESS 2	ESS 3	PSRF	ESS 1	ESS 2	ESS 3	PSRF	ESS 1	ESS 2	ESS 3	PSRF
Tree height	57	40	13	1.11 (1.33)	1063	1113	876	1.03 (1.10)	858	937	940	1.00 (1.01)
Tree length	55	34	7	1.17 (1.49)	1021	1150	907	1.03 (1.10)	860	1373	907	1.00 (1.01)
(log) Pop Size[1]	622	657	27	1.01 (1.02)	1098	833	751	1.00 (1.01)	230	1138	1261	1.02 (1.03)
CP1.alpha	7446	6992	7202	1.00 (1.00)	2705	2659	2484	1.00 (1.01)	3567	3377	3320	1.00 (1.00)
CP1.kappa	8620	8695	8640	1.00 (1.00)	4360	3796	3753	1.05 (1.07)	5204	5551	4912	1.00 (1.00)
CP2.kappa	6763	6876	7325	1.00 (1.00)	2674	2807	1983	1.00 (1.01)	3255	3761	3571	1.00 (1.00)
Mean rate	46	49	12	1.17 (1.49)	1058	1270	846	1.00 (1.00)	741	1396	1002	1.00 (1.01)
Coefficient of variation[2]	26	41	7	1.12 (1.28)	748	794	817	1.00 (1.00)	531	1147	1178	1.01 (1.02)
Covariance[3]	3370	3370	1315	1.00 (1.00)	5168	5209	5134	1.00 (1.00)	6148	6088	6124	1.00 (1.00)
Multivariate ESS/PSRF[4]	1299	1212	937	1.13	2018	2170	2034	1.02	1808	2632	2407	1.01

Table 3.2: Convergence diagnostics in phylogenetic space. I show convergence diagnostics specially tailored towards phylogenetic space, such as the approximate ESS of Lanfear et al. (2016) (for various metrics), the average univariate ESS for clade indicators[1], the clade switching score[2] and the average standard deviation in split/clade frequencies[3]. Tree ESSs computed using a sample of 1001 trees following the expression in (3.14).

	Poor			Default			STL		
	Chain 1	Chain 2	Chain 3	Chain 1	Chain 2	Chain 3	Chain 1	Chain 2	Chain 3
Tree ESS (KC, $\lambda = 0$)	33	38	39	743	1001	1001	523	734	465
Tree ESS (KC, $\lambda = 1/2$)	156	305	20	444	581	1001	292	231	445
Tree ESS (KC, $\lambda = 1$)	139	395	19	371	705	711	276	276	1001
Tree ESS (RF)	55	114	36	794	1001	801	695	624	787
Tree ESS (SP)	66	1001	14	403	659	659	184	155	552
Tree ESS (CD)	1001	431	87	1001	1001	1001	1001	1001	1001
Tree ESS (BS)	212	259	25	260	445	675	232	241	340
Mean clade ESS[1]	874	917	837	7321	7446	7420	7506	7477	7428
CSS[2]	0.16	0.18	0.16	0.87	0.88	0.88	0.88	0.88	0.9
ASDSF[3]	0.18	0.39	0.29	0.08	0.09	0.04	0.05	0.03	0.04



(a) $\lambda = 0$

(b) $\lambda = 1/2$

Figure 3.4: MDS projections for the full EBOV 1610 taxa data set (Kendall-Colijn distances). I sampled 200 equally spaced trees from each chain (1200 trees in total) and computed the Kendall-Colijn distance with $\lambda = 0$ and $\lambda = 1/2$. Colours relate to the seed used in the pseudo random number generator, ensuring each run starts from the same point. See Figure 3.3 for the projection of the same trees under the Robinson-Foulds metric (Robinson and Foulds, 1981).

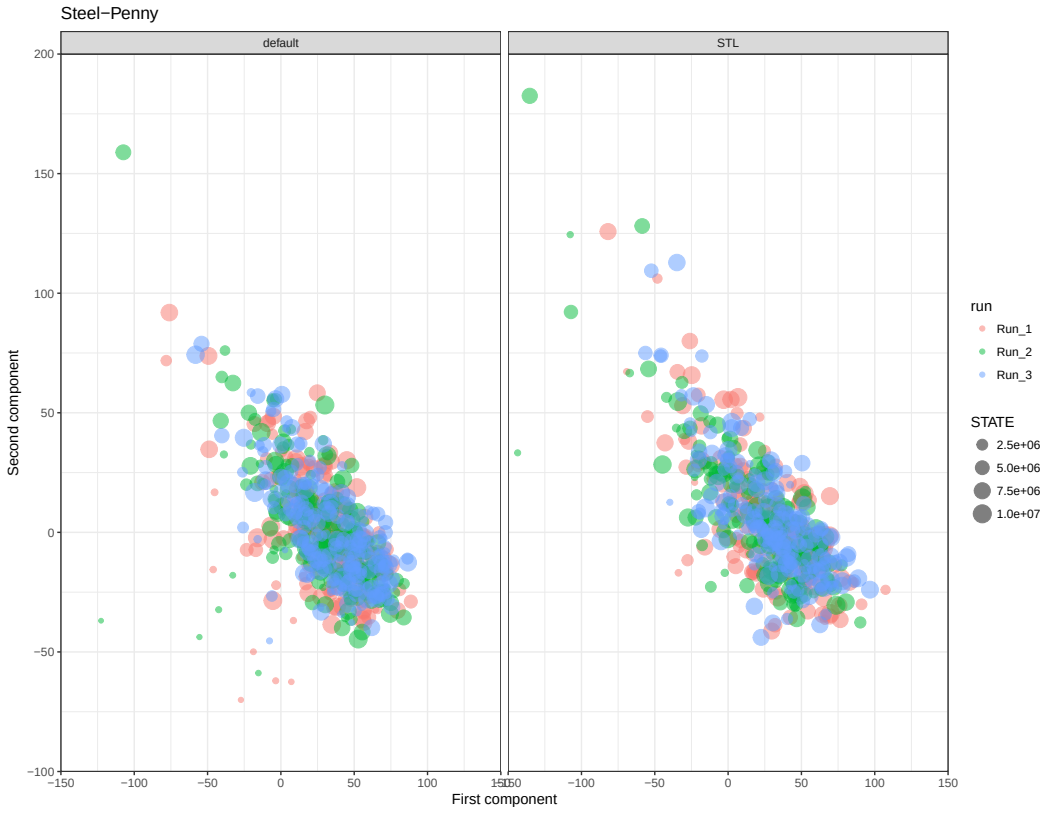


Figure 3.5: MDS projections for the full EBOV 1610 taxa data set (Steel-Penny distances). I sampled 200 equally spaced trees from each chain (1200 trees in total) and computed the Steel-Penny distance (Steel and Penny, 1993) for every pair of trees. Colours relate to the seed used in the pseudo random number generator, ensuring each run starts from the same point. See Figures 3.3 and 3.4 for projections of the same trees under different metrics.

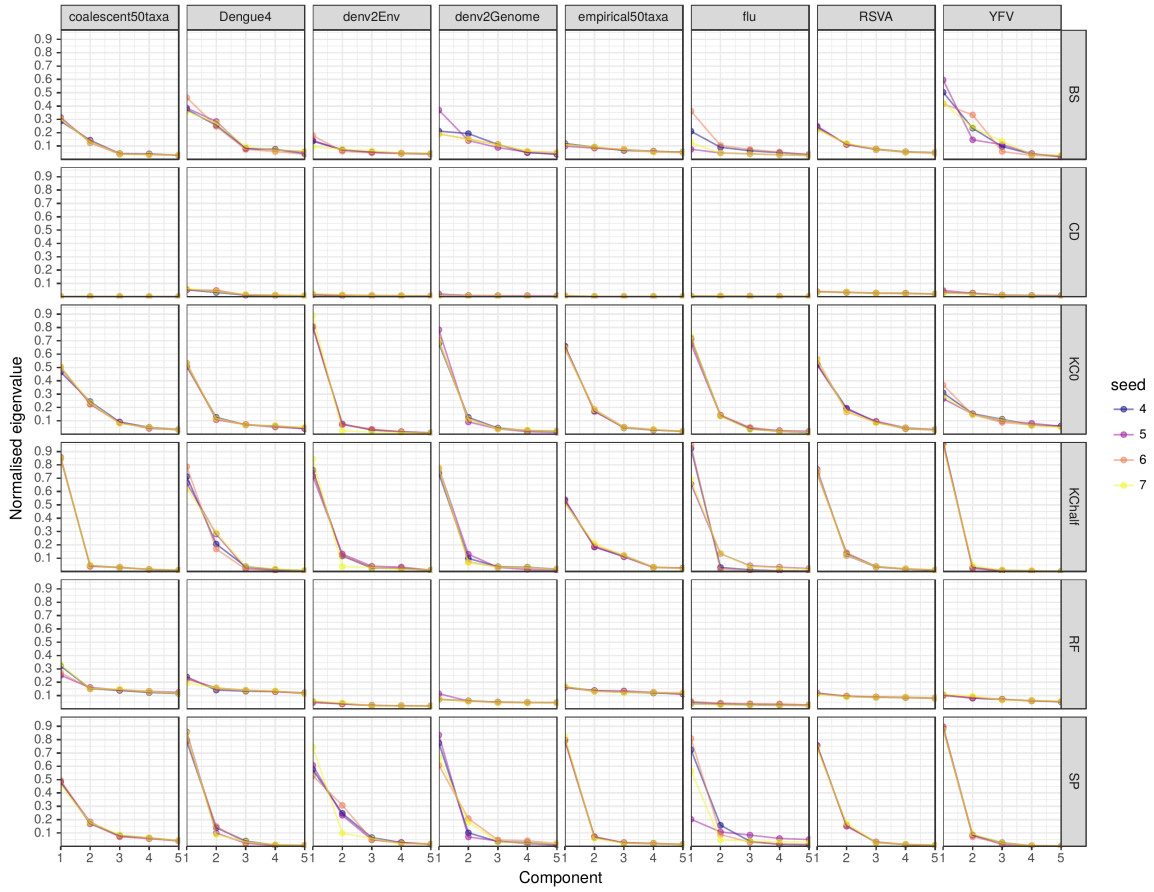


Figure 3.6: Scaled eigen values of phylogenetic space MDS. Vertical tiling shows the data set and horizontal ones show the phylogenetic metric used. Colours relate to the seed used in the pseudo random number generator, ensuring each run starts from the same point. "Elbow-shaped" plots indicate fewer components are needed to capture the variation in the data.

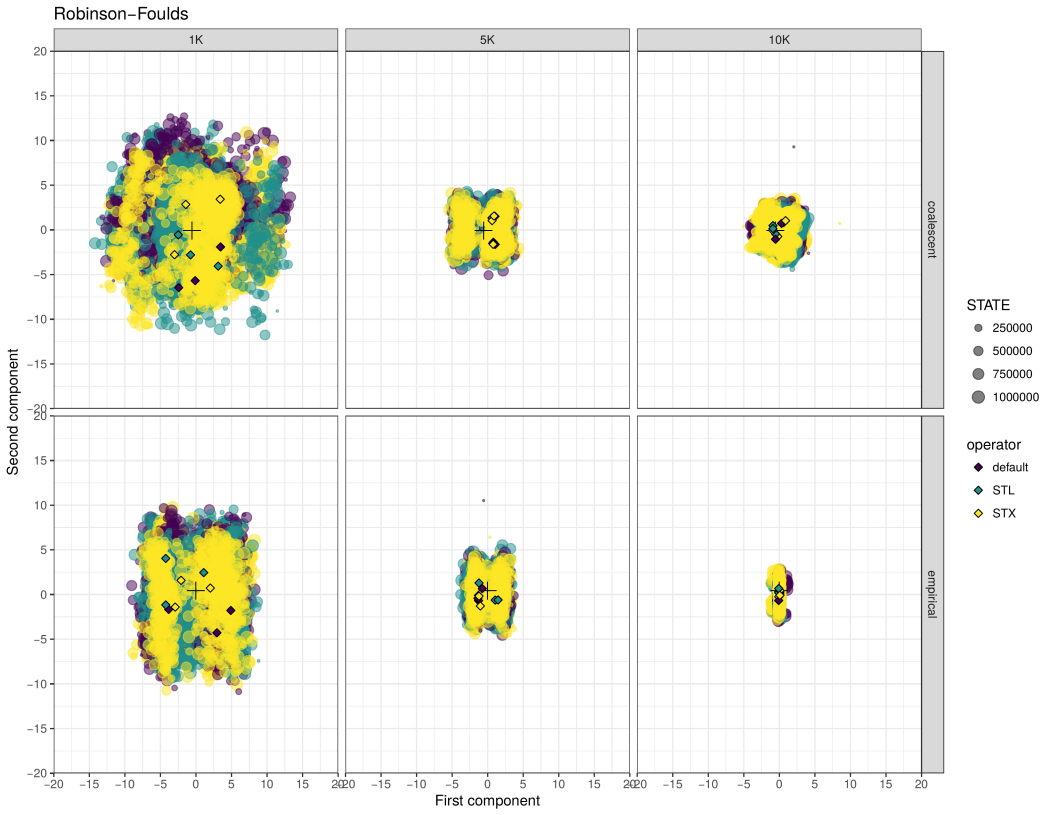


Figure 3.7: MDS projections for the simulated 50 taxa data set (Robinson-Foulds distances). I show three replicates per operator, 1000 trees in each replicate, computing the RF distance between all pairs of trees. The cross marks the true tree used to simulate the data. Colours pertain to the combination of MCMC transition kernels used and solid diamonds mark the maximum clade credibility (MCC) trees obtained from each run. Horizontal panels show the true tree: either drawn from the coalescent or extracted from a real world data set ("empirical"). Vertical panels show the number of sites in the simulated alignment (1000, 5000 or 10 000). Please see Section 3.4.1 for more details.

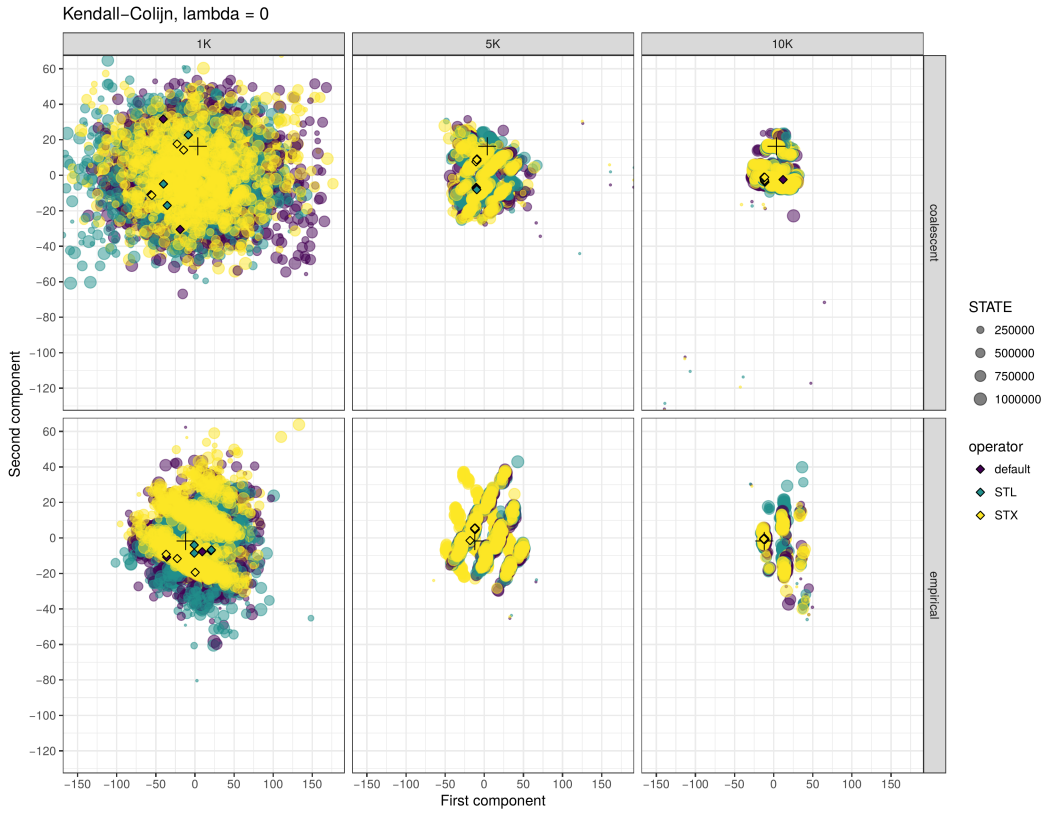


Figure 3.8: MDS projections for the simulated 50 taxa data set (Kendall-Colijn distances). I show three replicates per operator, 1000 trees in each replicate, computing the KC distance between all pairs of trees. The cross marks the true tree used to simulate the data. Colours pertain to the combination of MCMC transition kernels used and solid diamonds mark the maximum clade credibility (MCC) trees obtained from each run. Horizontal panels show the true tree: either drawn from the coalescent or extracted from a real world data set ("empirical"). Vertical panels show the number of sites in the simulated alignment (1000, 5000 or 10 000). See Figure 3.7 for a projection of the same trees under the RF metric.

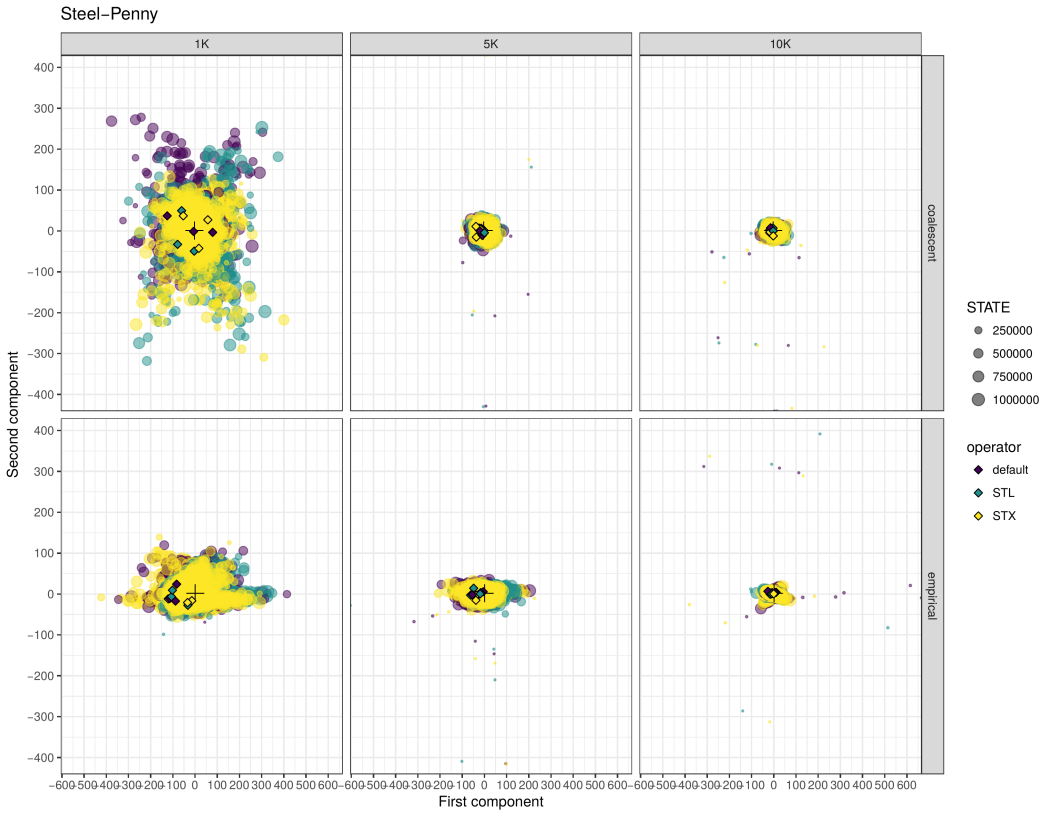


Figure 3.9: MDS projections for the simulated 50 taxa data set (Steel-Penny distances). I show three replicates per operator, 1000 trees in each replicate, computing the SP distance between all pairs of trees. Colours pertain to the combination of MCMC transition kernels used and solid diamonds mark the maximum clade credibility (MCC) trees obtained from each run. Horizontal panels show the true tree: either drawn from the coalescent or extracted from a real world data set ("empirical"). Vertical panels show the number of sites in the simulated alignment (1000, 5000 or 10 000). See Figure 3.7 for a projection of the same trees under the RF metric.

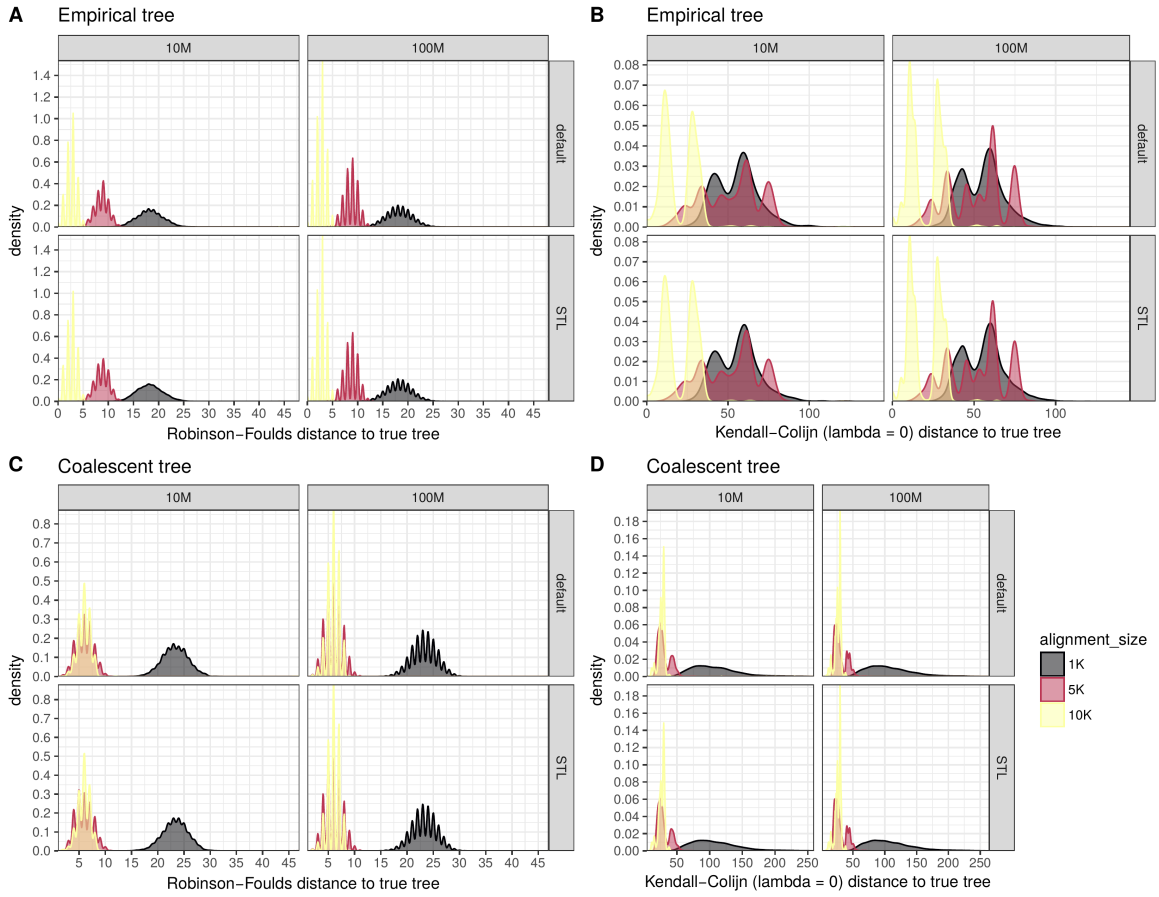


Figure 3.10: Characterisation of topological modes for a simulated example (50 taxa). I ran chains of 10 and 100 million iterations for three alignment sizes and two generating trees (see Chapter 2 for more details). Panels A and C show the Robinson-Foulds (Robinson and Foulds, 1981) distance to the tree, while panels B and D show the Kendall-Colijn (Kendall and Colijn, 2016) distance, with $\lambda = 0$ (topology only). Vertical panels show the chain length used, while horizontal panels display results obtained with different sets of transition kernels.

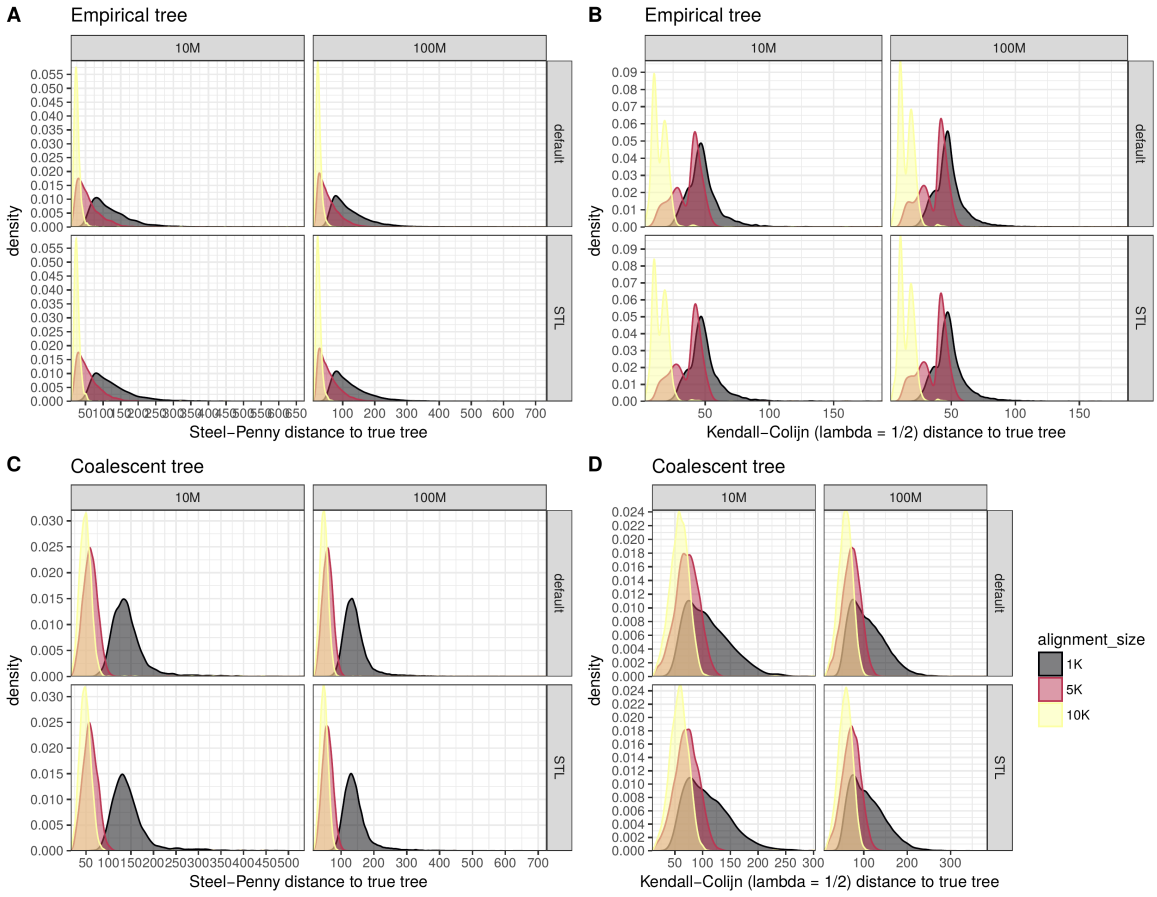


Figure 3.11: Characterisation of continuous phylogenetic space for a simulated example (50 taxa). I ran chains of 10 and 100 million iterations for three alignment sizes and two generating trees (see Chapter 2 for more details). Panels A and C show the Steel-Penny (Steel and Penny, 1993) distance to the tree, while panels B and D show the Kendall-Colijn (Kendall and Colijn, 2016) distance with $\lambda = 1/2$, which accounts for both topology and branch lengths. Vertical panels show the chain length used, while horizontal panels display results obtained with different sets of transition kernels.

Chapter 4

The epidemiological determinants of the 2013-2016 Ebola epidemic in West Africa

> Justice is like a snake; it only bites those who are barefoot.
>
> Eduardo Galeano (1940-2015) in *Is Justice just? (2009).*

In this chapter I detail my contributions to Dudas et al. (2017). I expand on the analysis presented in that paper by considering viral persistence times extracted from the phylogeographic analysis as a dependent variable. In addition, I consider spatial models, different error distributions and strategies to model comparison.

4.1 Introduction

By March 2014, the first cases of Ebola viral disease (EVD) were detected in Guinea, with subsequent epidemiological investigations suggesting the first cases occurred sometime around December 2013 (Baize et al., 2014). Until March 2016, a total of 28,616 confirmed, probable and suspected EVD cases have been reported in Guinea, Liberia and Sierra Leone, with 11,310 deaths (World Health Organization, 2016b). These figures make the 2013-2016 EVD epidemic in West Africa the worst in history, by far. Ebola virus (EBOV) was first detected in what is now the Democratic Republic of Congo (DRC), in 1976. Until the 2013-2016 epidemic, EBOV had been restricted to Middle Africa (Uganda, Sudan, DRC and Gabon) (Centers for Disease Control, 2015). Thus one of the main scientific questions emerging from the West African EVD epidemic was what factors led to such a high number of cases and geographic extent. Whilst many approaches are possible in tackling this question, in this chapter we will look at how virus genomes can be used to shed light into the ecological and epidemiological processes connected with the epidemic.

As the epidemic unfolded, various research groups started generating EBOV genomic sequences from clinical samples using high-throughput next-generation sequencing (NGS). In total, more than 1600 complete EBOV genomes were generated, resulting in a – temporally and spatially – dense sampling of over 5% of known cases. As argued by Holmes et al. (2016), this was an unprecedented scientific effort, and resulted in the best sampled genomic data set for an acute virus to date (see Chapter 6 for a more detailed discussion). Timely generation of high-quality sequences made it possible to gain insight into key aspects of the epidemic through the use of state-of-the-art phylogenetic methods (Dudas and Rambaut, 2014; Gire et al., 2014; Carroll et al., 2015; Park et al., 2015). Examples of such insights are the probable date of origin of the circulating strains (Gire et al., 2014; Park et al., 2015), the spatial spread of EBOV (Carroll et al., 2015; Dudas et al., 2017) as well as tracking transmission chains and understanding intra-host variability (Park et al., 2015). See Holmes et al.

(2016) for a comprehensive review of the studies generating EBOV genome sequences and the technologies employed.

While phylogenetic methods helped describe some aspects of the epidemic and EBOV evolution, the driving factors behind the epidemic in West Africa are largely unknown. To address this, Dudas et al. (2017) employed a sophisticated generalised linear model (GLM) framework to uncover the socio-economic, climatic and geographic factors driving EBOV spread in West Africa. Specifically, Dudas et al. (2017) put forth a phylogeographic model in which the transition (migration) rates between locations are modelled as linear combinations of a large set of predictors (Lemey et al., 2014). Examples of such predictors are the distance between locations, differences in population sizes, languages spoken, presence/absence of borders and climatic factors such as temperature and precipitation. This modelling strategy combines phylogenetic and as well as epidemiological information in a principled manner and allowed the authors to ascertain the relative importance of each predictor and thus uncover the driving factors behind EBOV spread.

In addition to understanding the factors associated with EBOV dispersal, it is also crucial to study the factors associated with local EBOV proliferation, measured by the number of detected cases. This question too involves assessing which covariates, amongst a large set, are strongly associated with the outcome of interest. In the remainder of this chapter I analyse the factors associated with local EBOV proliferation using a similar framework to that of Dudas et al. (2017). Importantly, I use information extracted from the phylogeographic models developed in that study as covariates in models for disease counts, thus improving on previous purely epidemiological models.

I detail the application of GLMs coupled with Bayesian stochastic search variable selection (BSSVS) to estimate parsimonious models that retain the most important from a large set of covariates. The fitted models are then used to predict the number of cases in countries not affected by the epidemic in order to provide insight into

why Ebola did not spread further. In addition, I apply the same framework to viral persistence data in order to study the driving factors behind viral maintenance.

4.2 Methods

Generalised linear models are a standard tool in Science and Engineering, providing a flexible way to investigate the relationship between an outcome variable (or variables) and a set of predictors or covariates. The remainder of this chapter assumes the reader is familiar with GLMs. A good introduction can be found in McCullagh and Nelder (1983).

In a scientific context, in addition to estimating coefficient values, the researcher is often interested in determining which covariates are more strongly associated with the outcome of interest. When P covariates are considered, this task usually entails selecting one amongst the 2^P possible models. The main idea of employing BSSVS with GLMs is to efficiently explore model space without however having to exhaustively visit all possible models. In this section I will lay out the approach of Kuo and Mallick (1998) to variable selection, the models developed and then proceed on to show how these models may be fitted to data and used for prediction.

4.2.1 Extended generalised linear model

Let $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}^\intercal$ be a set of N observations and let $\mathbf{X}$ be a $N \times P$ matrix of covariates measured without error. Finally, let $\boldsymbol{\delta} = \{\delta_1, \delta_2, \ldots, \delta_N\} \in [0,1]^P$ be set of indicator variables. The model we consider here is of the form

$$g(y_i) = \alpha + \sum_{k=1}^{P} \beta_k \delta_k X_{ki} + \epsilon_i,\ i = 1, 2, \ldots, N, \tag{4.1}$$

where $g(\cdot)$ is a *link function*, $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_P\}$ are the regression coefficients,

α is the intercept and $\epsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed (i.i.d) errors. One of the advantages of this formulation of the model is that one can readily interpret the parameters (and sub-models), in the following way: if $\delta_j = 1$, then the j-th predictor is included and when $\delta_j = 0$, the j-th predictor is omitted from the model. Here I will restrict attention to the case where the intercept (α) is always included. I also do not consider models with interactions, although these can be handled in a straightforward manner under the current framework (Kuo and Mallick, 1998, Eq. 1.2).

Another advantage is the ability to test for predictor relevance/importance by means of Bayes factors analytically (see next section for details). I note, however, that other approaches to variable selection are possible, see e.g. Mitchell and Beauchamp (1988); George and McCulloch (1993) and see O'Hara et al. (2009) for a review. This model can be very efficiently fitted to data using Markov chain Monte Carlo (MCMC), as I will detail in section 4.2.3.

The issue of shrinkage in regression, *i.e.* the reduction in the effects of sampling variation by obtaining a parsimonious representation of the data, is a long-standing one. Many prior formulations are possible when the goal is to perform variable selection (see Malsiner-Walli and Wagner (2016) for a review). The so-called spike-and-slab priors attempt to allow for shrinkage by either using an absolutely continuous distribution with a mode at zero or distributions with a point-mass at zero, called Dirac spikes. Here I will concentrate on constructions with Dirac spikes for $\boldsymbol{\beta}$. Specifically, for the continuous part I assign the coefficients a multivariate Gaussian prior, i.e., $\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, \tau \boldsymbol{I})$. where $\mathbf{0}$ is a P-dimensional vector of zeroes, $\boldsymbol{I}$ is the $P \times P$ identity matrix and τ is the variance of the coefficients. For simplicity, I will follow Lemey et al. (2014) and let $\tau = 4$.[1] See below for the prior on $\boldsymbol{\delta}$ that induces the spikes on $\boldsymbol{\theta} = \boldsymbol{\beta} \times \boldsymbol{\delta}$. As with the coefficients, various choices of priors for the intercept α are possible, $N(0, \tau)$ being a natural choice. Notice these prior choices are by no means restrictive; one may choose a different prior correlation

[1] Kuo and Mallick (1998) recommend $1/2 \leq \tau \leq 4$.

structure so as to incorporate problem-specific constraints or knowledge. Finally, the variance parameter σ^2 can be given an inverse-Gamma prior with parameters $\alpha = \beta = 10^{-3}$.

Note that these are general prior recommendations. Ideally, one should adapt their priors to suit the problem at hand. In the analyses presented in this chapter I will at times use different prior (and model) formulations in order to better address the scientific questions at hand. Importantly, this extended GLM framework is flexible and can be adapted to accommodate several distributions for $\boldsymbol{Y}$, as will be illustrated by the applications explored in this chapter.

Bayesian stochastic search variable selection

The idea behind Bayesian stochastic search variable selection (BSSVS) is to explore the space of 2^P possible models efficiently, visiting each sub-model proportional to its posterior probability conditional on the measured data. A central component to BSSVS is the prior $\pi(\boldsymbol{\delta})$. A natural choice of prior for $\boldsymbol{\delta}$ is $\Pr(\delta_j = 1) = p_j$, i.e., a Bernoulli prior on the indicator variables. This induces a binomial prior on $S = \sum_{k=1}^{P} \delta_k$. We can take advantage of this to construct priors for $\boldsymbol{\delta}$ that control how many predictors are included in the model, effectively controlling how parsimonious we want to be *a priori*. Let $p_j = q$, $j = 1, 2, \ldots, P$ and w be the probability of no predictors being included, that is, $w := \Pr(S = 0)$. I shall refer to w as the *stringency* of the prior on S. Then, it is straightforward to see that $w = (1 - q)^P$ and hence $q = 1 - w^{1/P}$. Here I will use $w = 1/2$. It is important to notice that when employing BSSVS, the researcher might want to place a prior directly on S For instance, Lemey et al. (2009) and Drummond and Suchard (2010), dealing with different applications, place a truncated Poisson prior on S.

As previously stated, one of the main advantages of the BSSVS approach is the ability to assess the relevance of covariates by computing Bayes factors. If we let $\hat{\delta}_j$ be an estimator of the posterior probability of δ_j, we can write the **Bayes factor**

BF_j for the j-th covariate as the ratio of posterior and prior odds, i.e.

$$\text{BF}_j = \frac{\hat{\delta}_j}{1-\hat{\delta}_j} / \frac{p_j}{1-p_j}, \tag{4.2}$$

$$= \frac{\hat{\delta}_j(1+w^{1/P})}{(1-\hat{\delta}_j)(1-w^{1/P})}. \tag{4.3}$$

A graphical representation of the relationship between $\hat{\delta}$, w and the Bayes factors is presented in Figure 4.1.

4.2.2 Modelling count data

Pertinent to the scientific questions addressed in this chapter is the issue of applying the model in 4.1 to count data. I now proceed to discuss some useful model extensions.

A natural choice in a GLM context is to assume that the data follow a Poisson distribution with mean λ, which implies a link function $g(\lambda) = \log(\lambda)$. Hence one can devise a model of the form

$$Y_i \sim \text{Poisson}(\lambda_i), \tag{4.4}$$

$$\log(\lambda_i) = \alpha + \sum_{k=1}^{P} \beta_k \delta_k X_{ki}. \tag{4.5}$$

This is a common and widely employed model for count data. It may be the case however that $Var(\boldsymbol{Y}) >> E[\boldsymbol{Y}]$, i.e., that there is considerable **overdispersion** in the data. Since under the Poisson distribution $E[\boldsymbol{Y}] = Var(\boldsymbol{Y}) = \lambda$, this choice of model may prove too restrictive. One way of accommodating overdispersion is the

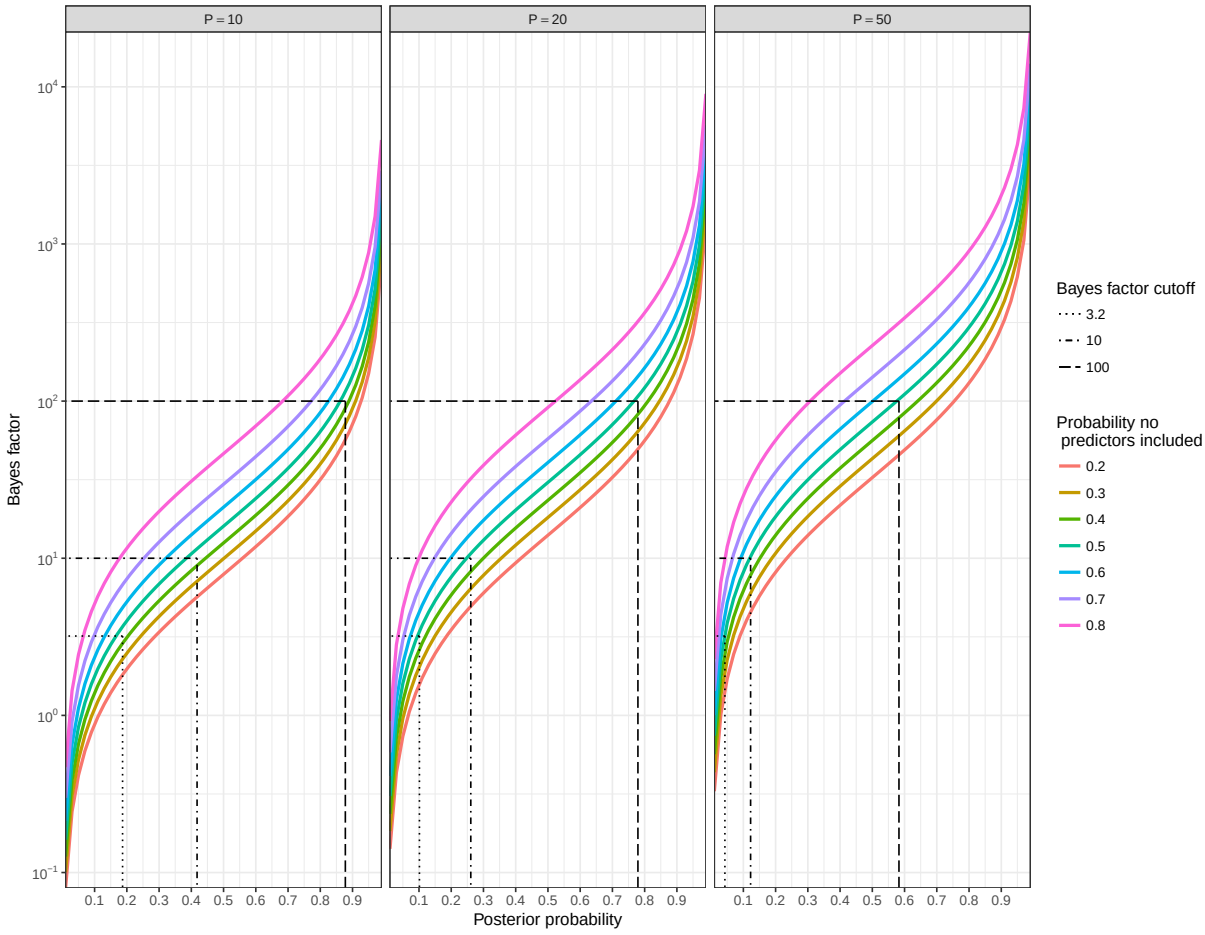


Figure 4.1: Visualising the relationship between prior stringency and Bayes factors for SSVS. Here I show the relationship between the posterior inclusion probability of a covariate and the Bayes factor for various levels of stringency. I use dotted (dashed) lines to show three common levels of "relevance" of the Bayes factor as presented in Kass and Raftery (1995) for the level of stringency adopted here ($w = 0.5$). The panels refer to the number of predictors (P) under consideration. Note how one needs to be careful about the calibration of w as P changes.

so-called "observation-level random effects" (ORLE) model, whereby an observation-specific error term is added (Hinde, 1982; Gelman and Hill, 2007; Harrison, 2014).

$$Y_i \sim \text{Poisson}(\lambda_i), \tag{4.6}$$

$$\log(\lambda_i) = \alpha + \sum_{k=1}^{P} \beta_k \delta_k X_{ki} + \epsilon_i, \tag{4.7}$$

$$\epsilon_i \sim N(0, \sigma^2), \tag{4.8}$$

leading to a non-standard distribution for $\boldsymbol{Y}$ that has $E[\boldsymbol{Y}] = \exp(\sigma^2/2)\lambda$ and $Var(\boldsymbol{Y}) = [\exp(2\sigma^2) - \exp(\sigma^2)]\lambda^2 + \exp(\sigma^2/2)\lambda$. When $\sigma^2 = 0$, we recover the model[2] in equation (4.4).

An alternative approach to modelling overdispersion is to assume the data come from a negative binomial distribution:

$$Y_i \sim \text{NegBin}(p_i, r), \tag{4.9}$$

$$p_i = \frac{r}{(r + \lambda_i)}, \tag{4.10}$$

$$\log(\lambda_i) = \alpha + \sum_{k=1}^{P} \beta_k \delta_k X_{ki}, \tag{4.11}$$

where r is the over-dispersion parameter. Under this model, $E[\boldsymbol{Y}] = \lambda$ and $Var(\boldsymbol{Y}) = \lambda + \lambda^2/r$. Note that r for this model plays a similar role as σ^2 for the model in 4.6. The choice of model can be done on a generative basis, i.e., relative to the data-generating process or following model fit criteria (see below).

If the counts take place in space, for instance if we are modelling disease counts, a model that takes spatial – and potentially temporal – information into account could be specified. Such a model is similar in structure to the ORLE model (4.6), with the addition of errors $\boldsymbol{\eta}$ that are not independent but assumed to follow some spatial process. A common and flexible choice is the conditional autorregressive

[2] I am keeping λ consistent across parametrisations for ease of comparison.

(CAR) model of Besag-York-Moillé (BYM) (Besag et al., 1991):

$$Y_i \sim \text{Poisson}(\lambda_i), \tag{4.12}$$

$$\log(\lambda_i) = \alpha + \sum_{k=1}^{P} \beta_k \delta_k X_{ki} + \epsilon_i + \eta_i, \tag{4.13}$$

where the errors are assumed to have conditional distributions of the form

$$\eta_i | \boldsymbol{\eta}_{-i} \sim \text{Normal}\left(\mu_i + \phi \sum_{i=1}^{N} \frac{W_{ij}}{D_{ii}} \left(\eta_j - \mu_j\right), \tau^2 D_{ii}^{-1}\right), \tag{4.14}$$

leading to the joint distribution

$$\boldsymbol{\eta} \sim \text{Normal}\left(\boldsymbol{\mu}, \tau^2 \left(\boldsymbol{D} - \phi \boldsymbol{W}\right)^{-1}\right), \tag{4.15}$$

where $\boldsymbol{W}$ is a spatial weight matrix, usually $W_{ij} = 1$ if i and j are spatial neighbours and $W_{ij} = 0$ otherwise, with $W_{ii} = 0$. A diagonal matrix $\boldsymbol{D}$ where $D_{ii} = \sum_{j=1}^{N} W_{ij}$, a vector of means $\boldsymbol{\mu}$, a correlation parameter ϕ and a scale parameter τ complete the model specification. A common simplification is to assume the spatial process has mean zero, i.e, $u_i = 0, \forall i$, since one is interested in describing the mean of the process using the covariates $\boldsymbol{X}$. Here I consider only the so-called proper models, i.e., models for which $|\phi| < 1$. Data are usually not very informative with respect to ϕ and τ^2, so it is customary to fix these parameters and then perform a sensitivity analysis to their fixed values. I have implemented both the fully Bayesian version and the fixed parameter version, this latter having the advantage of much faster computation times. For prior recommendations on the parameters of the models discussed in this section, I refer the viewer to Gelman (2006) for priors on the variance components and to Banerjee et al. (2003) (sec. 5.4.3) for CAR model parameters.

Finally, I remark that it is usual when modelling disease data to write the mean as a product of an exposure variable E_i and a relative risk, i.e., $\lambda_i = E_i e^{\psi_i}$. An example of exposure variable would be the population in a certain area.

4.2.3 Inference

The (joint) posterior distribution for the parameters in 4.1,

$$p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\delta}, \alpha, \sigma^2, \tau^2, ...\}$ stands for all parameters of interest, is not available analytically, but fortunately can be efficiently approximated using Markov chain Monte Carlo. I implemented all models discussed in this chapter using the language JAGS (Plummer et al., 2003). I run four independent chains for $M = 50000$ iterations. To draw inference about the model I then remove at least 10% of the chain as warm-up and then use the resulting sample for all subsequent computations (calculating summaries, prediction, etc.).

Let $\mathbf{C}_j = \{\delta_j^{(1)}, \delta_j^{(2)}, \ldots, \delta_j^{(M)}\} \in [0,1]^M$ be a MCMC sample for the j-th indicator. A natural choice of estimator for the posterior inclusion probability for the j-th covariate is $\hat{\delta_j} = (1/M)\sum_{k=1}^{M} \delta_j^{(k)}$. Convergence was determined by visually inspecting the trace of all values in the chain and by computing effective sample sizes (ESS). A chain was considered to have converged to the target distribution if all parameters had ESS$>$ 200. When β_j is close to zero the likelihood is nearly insensitive to whether $\delta_j = 0$ or $\delta_j = 1$, hence I monitor convergence of the product $\theta_j = \delta_j\beta_j$ instead, as advised by O'Hara et al. (2009). Also of interest is the conditional distribution $p(\beta_j|\delta_j = 1)$, which I use to report coefficient estimates.

Finally, from a Bayesian perspective prediction follows in a straightforward fashion from the posterior distribution:

$$p(\boldsymbol{Y}_{new}|\boldsymbol{Y}) = \int_{\Theta} p_1(\boldsymbol{Y}_{new}|\theta,\boldsymbol{Y})p_2(\theta|\boldsymbol{Y})d\theta.$$

This means that one can simultaneously estimate parameters θ and compute predictions $\boldsymbol{Y}_{new}$ within the same chain in MCMC.

4.2.4 Model comparison

Between the choice of error distribution and the structure of the varying intercepts (random effects), there are a host of possible models one could build. The issue of determining which model fits the data best is a central one in Statistics (Kass and Raftery, 1995; Vehtari et al., 2017). Here I will employ two information criteria to evaluate model fit, in addition to observed-versus-fitted analyses. The first information criterion I consider is the deviance information criterion (DIC) (Spiegelhalter et al., 2002). Let $D(\boldsymbol{\theta}) = -2\log(p(\boldsymbol{Y}|\boldsymbol{\theta}))$ be the model *deviance*. Then we define $\text{DIC} = E[D(\boldsymbol{\theta})] + pD$, where the expectation is taken w.r.t the posterior and pD is a model complexity penalty, calculated as in the commentary by Martyn Plummer in the discussion of Spiegelhalter et al. (2002).

A second information criterion I consider here is the widely applicable information criterion (WAIC) (Watanabe, 2010). WAIC is constructed as a fully Bayesian model fit quantity, and takes into account the out-of-sample predictive power of the model. WAIC is constructed as

$$\text{WAIC} = E[D(\boldsymbol{\theta})] + 2\sum_{i=1}^{N}\left(\log E[p(Y_i|\boldsymbol{\theta})] - E[\log p(Y_i|\boldsymbol{\theta})]\right),$$

and has the major advantage of not relying on the existence of a "true" model.

In addition to these univariate models of model fit, it is always prudent to assess model performance by evaluating its predictive power. In particular, predicted versus fitted plots are quite helpful in determining how well a model reproduces the data it was fitted to. By comparing prior and posterior predictive distributions for the data $\boldsymbol{Y}$ with its measured value one can gain insight not only into how well the model fits after data has been observed but also what sort of predictions the model generates using only the pre-data (i.e. prior) information.

4.3 Modelling Ebola in West Africa

In this section I describe the application of the modelling framework presented above to study the epidemic determinants of Ebola virus disease in West Africa. Following the study by Dudas et al. (2017) a rich set of socio-economic, climatic and genetic data was made available. This chapter is concerned with exploring these data to gain further insight into the factors driving EBOV proliferation and persistence.

4.3.1 Modelling case counts

The first question of interest is the relationship between the number of Ebola viral disease (EVD) cases and a plethora of socio-economic, climatic and genetic predictors. Following Dudas et al. (2017) I use administrative regions at the district (Sierra Leone), prefecture (Guinea) and county (Liberia) levels as the units of observation. EVD case numbers are reported by the WHO for every country division (region) at the appropriate administrative level, split by epidemiological week. I aggregate probable EVD cases and laboratory-confirmed cases in WHO reports[3]. This results in a data set with 81 locations, 63 of which reported one or more EVD cases. The predictors available for this analysis are listed in Table 4.1.

I expand the case count analyses we presented in Dudas et al. (2017) in two ways. Firstly, I include phylogenetic information in the form of the (average) number of predicted introductions, inferred from genetic data using a phylogeographic model. I use the posterior average number of introductions into a certain region inferred from phylogenetic data using either presence/absence of administrative boundaries (`VirIntroAdmin`) or distance between locations (`VirIntroDist`). Details of how these computations were done are provided in the supplementary information of Dudas et al. (2017).

[3]The main motivation for this approach is to – partially – accommodate subnotification.

Secondly, I consider a setting with 57 predictors whereas Dudas et al. (2017) only considered 13 predictors. This leads to a classical $N < P$ situation, i.e, a situation where one has as many or more covariates/predictors as one has data points. My goals are: (i) to investigate the association of predicted introductions and case numbers; (ii) to investigate some predictors not included in the original analysis and (iii) to predict how many cases would have occurred at the locations which reported zero EVD cases, i.e, the potential size of the epidemic in each location.

Pursuing goal (i) would allow me to assess the hypothesis that the EVD epidemic in West Africa was mainly driven by re-introductions whereas goal (ii) could not only lead to new insight into the driving factors behind the epidemic but also would allow me to investigate the power of BSSVS in a $N < P$ scenario. Goal (iii) is important to assess the risk of new epidemics.

Table 4.1: Covariates considered for the case and persistence data modelling. All continuous predictors were standardised by subtracting the mean and dividing by the sample standard deviation.

Predictor type	Abbreviation	Predictor description	Included[1]
Demographic	PopDens	Population density (inhabitants/Km^2)	yes
Demographic	TTXk	Estimated mean travel time in minutes to reach the nearest major settlement of at least X $\times$ 10,000 people, for X = 50, 100 and 500.	yes
Demographic	GrEconMN	Mean gridded economic output	yes
Demographic	GrEconMIN	Minimum gridded economic output	yes
Demographic	GrEconMAX	Maximum gridded economic output	yes
Demographic	GrEconSTD	Standard deviations of gridded economic output	yes
Demographic	LangX	Percentage of the population speaking language X, for $X = 1, 2, \ldots, 17$	no
Climatic	AltMN	Mean altitude (elevation above sea level)	yes
Climatic	TempMN	Mean annual temperature	yes
Climatic	TempSSMN	Mean index of temperature seasonality	yes
Climatic	PrecipMN	Mean annual precipitation annual mean	yes
Climatic	PrecipSSMN	Mean annual seasonality in precipitation	yes
Climatic	PrecipX	Mean precipitation for month X, $X = 1, 2, \ldots, 12$	no
Climatic	HumX	Mean humidity for month X, $X = 1, 2, \ldots, 12$	no
Second order phylogenetic	VirIntroAdmin	Mean number of viral introductions predicted using administrative borders	no
Second order phylogenetic	VirIntroDist	Mean number of viral introductions predicted using distances between locations	no
Genetic	NumSeq	number of EBOV complete genome sequences collected	no

[1]Whether the predictor was included in the original analysis in Dudas et al. (2017).

EBOV persistence times

A second outcome of interest is the persistence of EBOV and its determinants. Combining genetic data and a phylogeographic model, Dudas et al. (2017) computed the numbers of viral introductions from and to each location (where sequences were sampled) and then estimated the persistence times (in days) for each location. As in Dudas et al. (2017), I define a cluster as a group of sequenced cases in a region that can be traced to a single introduction and then define persistence as the time from the MRCA and the last sampled sequence in the cluster. For this data set, $N = 56$ data points were available, corresponding to the 56 (admin 2 level) locations represented in the data analysed by Dudas et al. (2017).

As with the analysis of case counts, my main goal is to ascertain which of the available predictors is strongly associated with viral persistence in a given region. Since persistence times are a continuous, strictly positive outcome, I chose the log-normal and Gamma family of distributions to model its variation. The hierarchical model presented in 4.12, for instance, can be extended to any exponential family distribution, as shown in (Banerjee et al., 2003, Sec. 5.5). For these analyses I included the number of sequences per location (`NumSeq`) and the (standardised) population size (`PopSize`) as predictors. This is important as the number of sequences sampled in a given location control the "sample size" from which persistence times are computed. The more sequences sampled in a region, the more tips will be available to compute persistence times. In the same fashion as the analyses of case counts, I also consider an additional situation where a larger number of predictors are available ($P = 58$, $N = 56$).

4.4 Results and discussion

In what follows I present the results of fitting 10 models to each data set in fully Bayesian fashion from which I compute Bayes factors and extract (posterior)

predictions. I employ the Bayes factors from SSVS to tease apart which factors are associated with the outcome of interest and use posterior predictions for the regions with no reported cases (case data) or no sequences (persistence data) as a way o gaining insight into a region's potential for sustaining transmission chains in both size and duration.

For the EVD case data, Table 4.2 shows that a Poisson model with (unstructured) observation-level "random effects" (OLRE, model 2) showed best performance in terms of fit and hence I chose its SVSS counterpart, model 7, as the basis for computing Bayes factors for the predictors.

Table 4.2: Modelling results for EVD case data.

Model	Distribution	Intercept	Variable inclusion	DIC	WAIC	RMSE[1]	Obs. in CI[2]
0	Poisson	Single intercept	Full	5204	8717.3	1.83×10^2	15
1	Negative Binomial	Single intercept	Full	738.8	731.4	2.15×10^4	62
2	Poisson	OLRE	Full	489	**476.4**	**0.73**	63
3	Poisson	OLRE/spatial[3]	Full	481.5	478.1	0.84	63
4	Negative Binomial	OLRE/spatial[4]	Full	737.9	731.4	2.14×10^4	63
5	Poisson	Single intercept	SSVS	5181	8492.9	1.85×10^2	14
6	Negative Binomial	Single intercept	SSVS	763.6	747.2	7.38×10^2	63
7	Poisson	OLRE	SSVS	490.9	478.3	0.76	63
8	Poisson	OLRE/spatial[3]	SSVS	**475.8**	477.4	1.14	63
9	Negative Binomial	OLRE/spatial[4]	SSVS	758.2	747.6	6.49×10^2	63

[1]Root mean squared error, $\sqrt{\sum_{i=1}^{N}\left(Y_i-\hat{Y}_i\right)^2/N}$.

[2]Number of observations – out of $N = 63$ – within the predicted 95% credibility interval (CI).

[3]This is the BYM model as in 4.12.

[4]These models do not have the ORLE component.

In a similar fashion, in Table 4.3 I present the results of fitting 10 models to the persistence time data described above. Results point model 2 (log-normal OLRE) as the best performing model. However, the SSVS counterpart to model 2, model 7, shows worse performance when compared with model 8, which includes a spatial component. I thus decide to use model 8 for all further investigations.

Table 4.3: Modelling results for EVD persistence times.

Model	Distribution	Intercept	Variable inclusion	DIC	WAIC	RMSE[1]	Obs. in CI[2]
0	Log-normal	Single intercept	Full	495.9	502.7	12932.85	55
1	Gamma	Single intercept	Full	507.1	489.6	14739.0	54
2	Log-normal	OLRE	Full	**428.5**	**329.1**	142.6	56
3	Log-normal	OLRE/spatial[3]	Full	508.8	362.9	198.3	56
4	Gamma	OLRE/spatial[4]	Full	483.8	489	9889.6	56
5	Log-normal	Single intercept	SSVS	496.7	494.5	19838.0	54
6	Gamma	Single intercept	SSVS	487.8	486.1	19480.4	53
7	Log-normal	OLRE	SSVS	496.8	377.4	360.5	56
8	Log-normal	OLRE/spatial[3]	SSVS	452.2	330.2	**51.6**	56
9	Gamma	OLRE/spatial[4]	SSVS	491.3	488.6	14554.1	55

[1]Root mean squared error, $\sqrt{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 / N}$.
[2]Number of observations – out of $N = 56$ – within the predicted 95% credibility interval (CI).
[3]This is a modified BYM model (4.12) to accommodate different error distributions.
[4]These models do not have the ORLE component.

Epidemiological findings

I use the models to answer to major questions emerging from the 2013-2016 West African EVD epidemic: (i) what drove the local proliferation of transmission chains (cases) and (ii) which factors were associated with viral persistence in a given region.

What are the driving factors behind EBOV proliferation and persistence?

Using model 7 in Table 4.2, I computed the Bayes factors for all $P = 15$ predictors and keep those that show "strong evidence" according to Kass and Raftery (1995) (BF > 3), presented in Table 4.4. In agreement with the previous findings of Dudas et al. (2017)[4], the predictors strongly associated with case

[4]See Table 2 therein for comparison. Note that the analysis presented here is related to, but distinct from the one in Dudas et al. (2017).

counts are environmental factors such as temperature seasonality, precipitation and temperature. Additionally, travel time to the nearest 50 and 100 thousand inhabitants settlement, which are socio-economic variables that reflect a region's connectedness, also had higher Bayes factors. This result is consistent with Suchar et al. (2018), who also found that infrastructure variables were associated with epidemic potential using exploratory analysis. It should be noted that the results presented here differ from the ones in Dudas et al. (2017) in that the authors chose to include population sizes as a covariate, whereas I opt to use population sizes as an exposure variable and thus model relative rates.

Importantly, the genetic covariates relating to the numbers of viral introductions (`VirIntroAdmin` and `VirIntroDist`) were not strongly associated with case counts. This is a puzzling result, since all the evidence points towards the fact that the 2013-2016 EVD epidemic was mainly driven by migration of infected individuals (i.e., a "sparks and fires" model). It is possible that these variables were in fact poor proxies for the actual numbers of viral introductions into a given region. I deliberately chose these two variables because they did not include any "gravity" information, that is information about population sizes. This was to avoid correlation between the covariates and the exposure variable. It remains to be seen whether a better measure of viral introductions would help explain case counts. Another possibility is a complex interplay between the number of introductions a region receives and the local conditions (environmental, socio-economic, cultural) that cannot be captured by a linear relationship.

Table 4.4: Variable selection for the EVD case data. I present the results from model 7 (see Table 4.2).

Predictor[1]	Coefficient[2]	95% CI[3]	Inclusion[4]	BF[5]
TempSSMN	-1.52	-2.17, -0.70	0.82	99.6
PrecipMN	1.15	0.35, 1.91	0.35	11.6
TT50K	-0.71	-1.21, -0.27	0.31	9.4
TempMN	0.63	0.18, 1.08	0.20	5.4
TT100K	-0.59	-1.11, 0.64	0.17	4.2

[1]Predictors with Bayes factor >3.
[2]Posterior mean of the coefficient.
[3]95% posterior credible interval (CI).
[4]Posterior inclusion probability for the predictor.
[5]Bayes factor (BF).

I extend the analyses in Dudas et al. (2017) by considering a much larger set of predictors ($P = 57$, description in Table 4.1), taking advantage of the SSVS framework. Results in Table 4.5 are wholly consistent with the previous results, with most of the selected covariates being environmental variables related to precipitation, humidity and temperature. An important exception is the inclusion of the variable `Lang4` which measures the percent of speakers of a language spoken mostly in the north west of Guinea[5], a region with a relatively low number of observed cases. This variable was not considered by Dudas et al. (2017) and its inclusion with high probability indicates that cultural factors might shape local transmissibility conditions (see Alexander et al. (2015) and references therein). An alternative view is that the languages spoken in a region are a proxy for the spatial coherence of these regions, but this hypothesis does not explain why the language predictor would still be well supported after accounting for spatial dependence. Further investigation is needed before more elaborate hypotheses can be formulated.

[5]Unfortunately the data coding prevents me from knowing the exact language corresponding to Lang4.

Table 4.5: Variable selection for the EVD case data with $P = 57$ predictors. As with the results in Table 4.4, I present the results from model 7 and include all covariates with Bayes factor larger than 3.

Predictor[1]	Coefficient[2]	95% CI[3]	Inclusion[4]	BF[5]
Lang4	-1.12	-1.66, -0.61	0.66	153.9
Pre10MN	1.05	0.63, 1.58	0.53	91.2
Pre08MN	1.45	0.70, 2.12	0.27	30.2
pre04MN	1.54	0.72, 2.04	0.23	23.6
Hum08MN	1.04	0.35, 1.60	0.14	12.8
Pre09MN	1.04	0.43, 2.02	0.12	11.1
Hum04MN	-1.43	-2.94, -0.43	0.08	7.0
Pre11MN	0.90	0.15, 1.57	0.04	3.0
PrecMN	1.46	0.37, 2.20	0.04	3.0
Hum05MN	-1.14	-2.83, -0.11	0.04	3.0
TempMN	-0.64	-1.09, -0.23	0.03	3.0
PrecssMN	-1.26	-2.03, 0.16	0.03	3.0

[1]Predictors with Bayes factor >3.
[2]Posterior mean of the coefficient.
[3]95% posterior credible interval (CI).
[4]Posterior inclusion probability for the predictor.
[5]Bayes factor (BF).

When looking at persistence times and considering $P = 15$ predictors, the only predictor with a strong association with the outcome was the number of sequences in a region (Table 4.6), which was included as a control variable. Because of how the persistence times are computed, one expects them to be positively correlated with the number of sequences, hence this result is unsurprising. On the other hand, lack of association with any of the other of variables is itself informative insofar as it suggests that we might not be able to explain viral persistence by considering solely local factors.

Table 4.6: Variable selection for persistence time data. I present the results from model 8 (see Table 4.3).

Predictor[1]	Coefficient[2]	95% CI[3]	Inclusion[4]	BF[5]
Sequence count	0.19	0.08, 0.33	0.21	5.77

[1]Predictors with Bayes factor >3.
[2]Posterior mean of the coefficient.
[3]95% posterior credible interval (CI).
[4]Posterior inclusion probability for the predictor.
[5]Bayes factor (BF).

I refine my analysis by considering a larger set of $P = 58$ covariates and find that, in addition to sequence counts as before, only one more variable shows substantial association with persistence times (Table 4.7). The variable `Precip01`, which measures the average precipitation in January in a given region, was included with high probability. This result is hard to explain since there is no obvious link between the amount of rain in a particular month and favourable conditions for viral lineage persistence. Interestingly, (standardised) population sizes failed to attain a high inclusion probability in any of the SSVS models, suggesting that, somewhat counterintuitively, how big the population is in a given area does not significantly impact its suitability for persisting lineages.

Table 4.7: Variable selection for persistence time data with $P = 58$ predictors. I present the results from model 8 (see Table 4.3).

Predictor[1]	Coefficient[2]	95% CI[3]	Inclusion[4]	BF[5]
Precip01	-0.35	-0.51, -0.16	0.11	10.0
Sequence count	0.20	0.07, 0.31	0.06	5.2

[1]Predictors with Bayes factor >3.
[2]Posterior mean of the coefficient.
[3]95% posterior credible interval (CI).
[4]Posterior inclusion probability for the predictor.
[5]Bayes factor (BF).

Why did the epidemic not spread further?

Whether an epidemic develops in a region depends on multiple factors, mainly (i) whether the virus is introduced with high frequency and (ii) the local conditions are favourable for maintenance of sustained transmission chains. While this chapter is mainly concerned with (ii), a complete picture of the epidemic can only be painted by also including (i). While I tried to address (i) by including the variables `VirIntroAdmin` and `VirIntroDist`, I now focus on (ii) with the hope of gaining insight into why some regions did not experience EVD cases.

Figure 4.2 shows the observed and predicted numbers of EVD cases and also the incidence per 100, 000 inhabitants. This is important because even though the absolute number of cases is an important epidemiological variable, incidence rates provide a relative (scaled) measure of the disease burden in a given region. Using a Poisson model with OLRE and SSVS (model 7, Table 4.4), I predict that Cavally and Tonkpi (in the Ivory Coast) for instance would have experienced 470 (1, 2547) and 1225 (0, 8854) cases, respectively. Despite these numbers being reasonably high compared to the overall average of 387 (95% credible interval: 2, 2206), they would mean incidences of 111 and 129 cases per 100, 000 inhabitants, which are not substantially high compared to the average of 97 (2, 408) cases/100K. These are predictions of the epidemic potential, however, as these regions did not report any cases. This suggests that while local conditions were conducive to the development of epidemics, the number of introductions into these locations was not enough to lead to successful establishment of transmission chains.

When considering predictions for persistence times, I show results for models 2 and 8 (see Table 4.3 for details). In order to identify areas with a particularly higher suitability for persistence, I select the locations with predicted values in excess of the 75% percentile of the predictive distribution of persistence times. Predictions from model 2 showed that Kenieba (Mali), Saraya, Tambacounda, Velingara (Senegal) were above the threshold. Using a log-normal distribution with

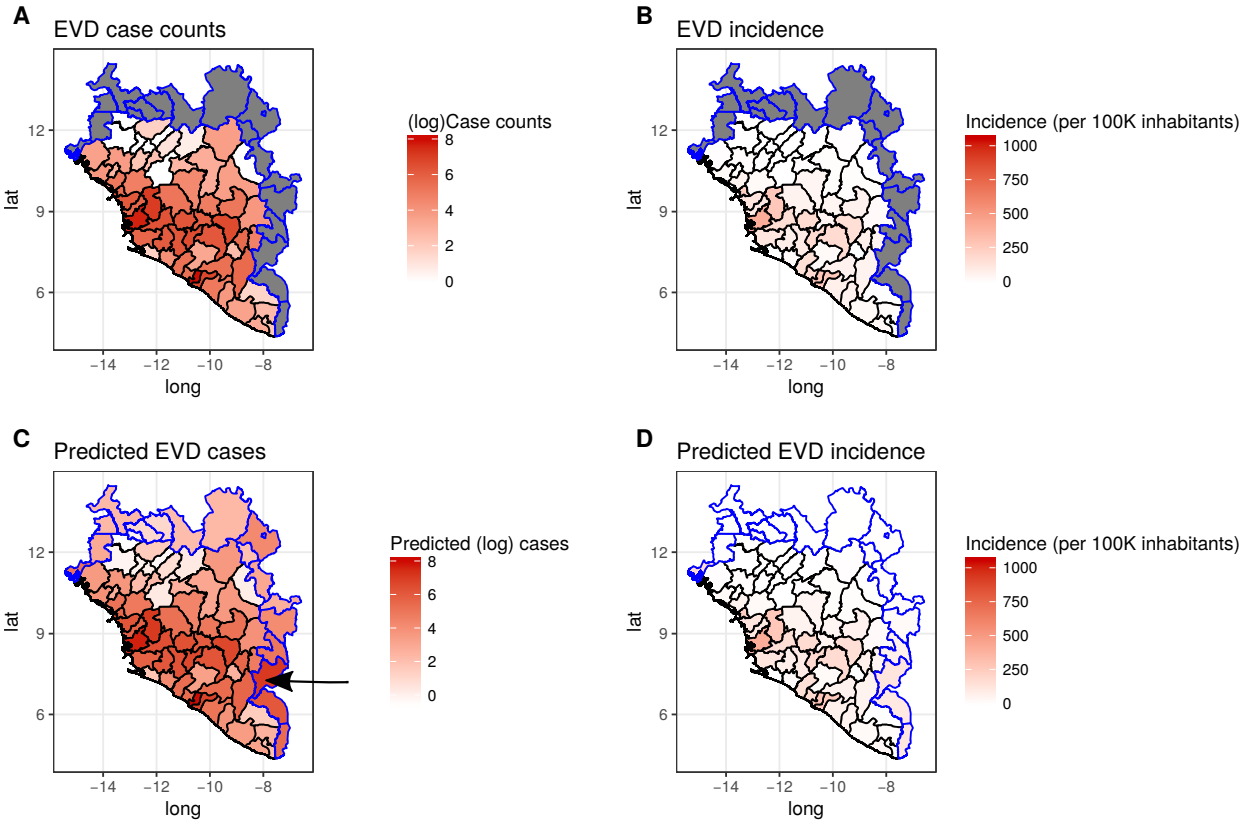


Figure 4.2: Posterior means for the numbers of EVD cases and incidence. Shaded areas show locations for which no cases were reported. Results from the Poisson model with observation-level varying intercepts ("OLRE") and SSVS (model 7) show reasonable in-sample predictive ability. Regions such as Tonkpi (indicated with an arrow) were predicted to have moderate EVD incidence despite being predicted to have higher than average numbers of EVD cases.

OLRE and spatial structure (model 8), the locations Bafing, Kabadougou, San Pedro (Ivory Coast), Kedougou, Salemata, Velingara (Senegal), Tombali, Gabu (Guinea-Bissau), Kenieba, Kati, Kangaba and Yanfolila (Mali) were all predicted to be above the 75% percentile. As with the case counts, the predictions seemed to reflect the pattern of the observed data quite well (Figure 4.3). Interestingly, none of these locations were predicted to have high numbers of cases. This result is consistent with the findings in Tables 4.4, 4.5, 4.6 and 4.7, that showed almost no overlap between the factors associated with EVD case numbers and persistence times.

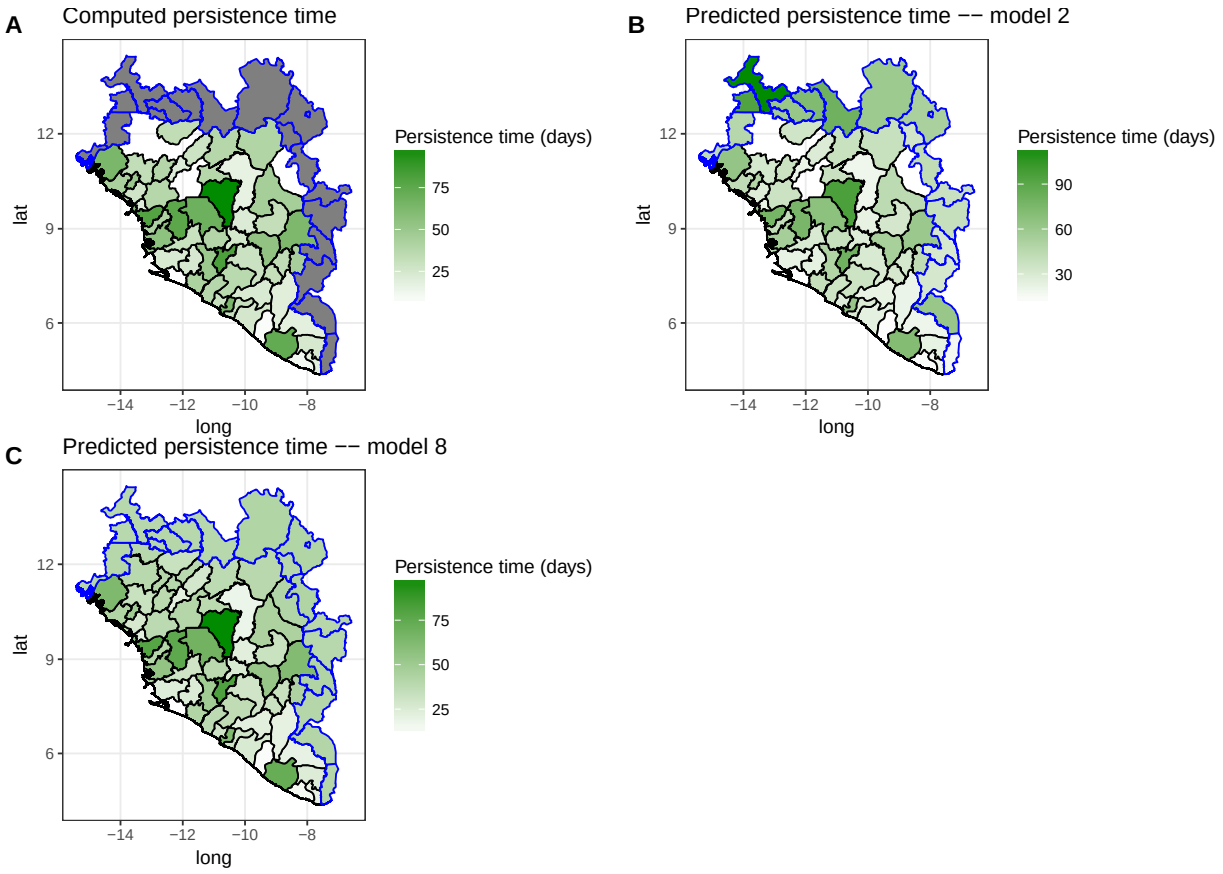


Figure 4.3: Posterior means of the computed by Dudas et al. (2017) and predicted EBOV persistence times (in days). I present predictions from models 2 and 8 (see Table 4.3).

In short, the lack of overlap between the regions that were predicted to have higher numbers of cases and those predicted to have higher persistence times might help

explain why the epidemic did not spread into these regions. The role of border closure and other control measures should not be disregarded, however. See Dellicour et al. (2017) for an interesting follow-up paper employing phylodynamic methods to address which interventions would have been effective in curbing the epidemic in West Africa. The complex interplay between migration of infected individuals between locations and the local "suitability" of each region is evidenced by the lack of clear spatial structure (see discussion on spatial models below). Complex patterns of migration of infected individuals – over long distances for instance – and border closure mean that two areas that are geographic neighbours would have very different epidemic dynamics, hence leading to a situation with low spatial autocorrelation.

Hence, in summary:

- Some regions that report no EVD cases were predicted to have high epidemic potential;
- Several regions in Mali and Senegal were predicted to have higher suitability for viral persistence;
- The lack of overlap between areas with high predicted numbers of cases and persistence times can explain why the epidemic did not spread further.

Methodological findings

I now focus on the technical aspects of the modelling effort presented in this chapter. I would like to point out that while I lay out a complete framework for modelling epidemiological data in this chapter, I make no claims of originality. See the frameworks proposed by e.g. Scheel et al. (2013) and Boehm Vock et al. (2015), who touch on the same ideas explored in this chapter: generalised linear models, spatial structures and variable selection.

A first question one might ask is whether including an explicit spatial component

improves model fit and (predictive) performance. To assess the amount of variation that is structured spatially, one can compute $\gamma = \text{sd}(\boldsymbol{\eta})/\left(\text{sd}(\boldsymbol{\eta}) + \text{sd}(\boldsymbol{\epsilon})\right)$. The closer γ is to 1 the more spatial structure there is in the data. For all of the models considered here including both data sets (cases and persistence) γ was estimated in the range $0.40 - 0.60$ with substantially wide credibility intervals. This points decisively in the direction of the **absence** of appreciable (residual) spatial structure. It is possible that a better calibrated spatial model would be able to capture spatial variation and lead to better performance. One such model is the one developed by Riebler et al. (2016) where the authors propose a scaled spatial component that improves model fit by generating better calibrated data *a priori*.

In keeping with these results, I find that models that include an spatial component perform no substantially better than models that include an unstructured, observation-level error term (see RMSE in Tables 4.2 and 4.3). In general, Poisson models with OLRE, whether an explicit spatial component was present or not (i.e., models 2, 3 7 and 8) presented better fit and predictive performance. From a methodological standpoint, one could argue that adding a spatially-varying error term might lead to overfitting when no extra-Poisson variation exists, but this has been shown to be a minor concern unless there is very strong spatial autocorrelation (Latouche et al., 2007). An attentive reader will notice I do not include an ORLE term in the negative binomial (cases) and Gamma (persistence) models. This is to avoid parameter identifiability issues between the ORLE variance σ^2 and the overdispersion (r) and shape (k) parameters respectively.

A second question that may arise is whether employing BSSVS to select parsimonious models significantly reduces model fit and predictive performance. Overall, results show that models where all covariates are included ("Full") had better fit and predictive performance. This pattern was consistent between the models for cases and persistence times – hence 10 pairs of model comparisons. However, these differences were not particularly marked. In addition, while the analyses with many predictors ($P \geq N$) helped refine our understanding of the factors associated with

both epidemic potential (cases) and viral persistence, I observed no gain in predictive performance for any of the 10 models[6] considered. This result is not entirely surprising, since including more predictors is bound to increase unconditional model fit by increasing the amount of data variation explained by the model. However, properly calibrated information criteria should penalise model complexity accordingly. In particular, in a Bayesian setting including more predictors also means incorporating uncertainty about the extra parameters. Moreover, while in-sample predictive ability should also increase with the number of predictors, ideally predictive ability should be judged mainly with regard to out-of-sample predictive performance. Future work will include assessing model fit using (Bayesian) leave-one-out (LOO) cross validation (Vehtari et al., 2017).

The BSSVS framework employed here has two major strengths. On one hand it allows one to analytically compute Bayes factors for predictors and thus formally assess the relevance of the association between predictor and outcome of interest. On the other, prior modelling is straightforward insofar as stringency/parsimony is concerned (see Figure 4.1). I note that whilst we make the assumptions of (prior) independence and exchangeability in order to greatly simplify calculations and implementation, it is possible to explicitly include dependencies between predictors (Chipman, 1996). Unfortunately, and perhaps witness to the plethora of variable selection methods currently available, BSSVS also has major flaws that limit its applicability. One such flaw is almost obvious to the trained eye: because of multicolinearity between covariates, we expect models that include either variable from a pair of highly correlated covariates to be about equally as probable. This rationale can be extended to see that several models amongst the 2^P possible will have virtually the same posterior mass. One could attempt to circumvent this fundamental multimodality by careful prior modelling, but this is not pursued further in this chapter. Another possibility, also not explored here but of interest for future research is to replace the discrete BSSVS variable selection procedure by a

[6]For each data set, there were five models with SSVS.

continuous, cross-validation-based method such as the one developed by Piironen and Vehtari (2017). Such an approach, while forfeiting the benefits of explicit Bayes factors for evaluating predictor support, also avoids the inherent combinatorial multimodality induced by the BSSVS formulation.

Finally, I would like to add a note about prior calibration for the class of models considered here. It is clear from the prior predictive analyses that the usually recommended priors on the coefficients and other parameters induce a prior on the data $\boldsymbol{Y}$ that is miscalibrated. For all the models considered for both data sets, sampling from $\pi(\boldsymbol{\theta})$ led to an induced distribution on $\boldsymbol{Y}$ that was more than 10 orders of magnitude off (Figure 4.4). Unfortunately, a complete prior calibration study is outside the scope of this chapter, but remains an open avenue for future research.

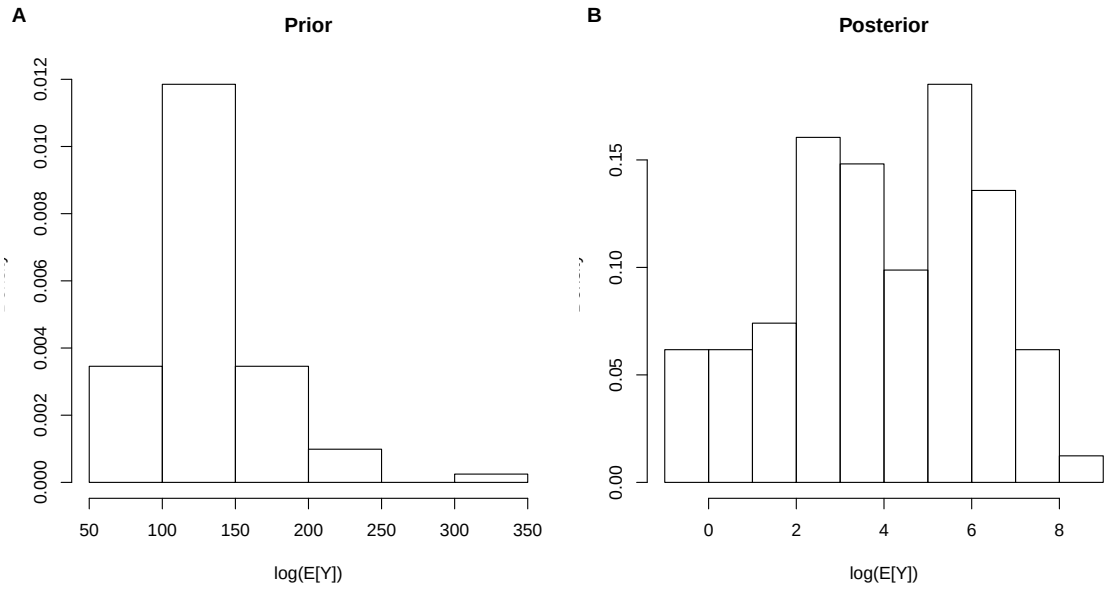


Figure 4.4: Prior and posterior predictive distributions of case counts, model 8 I present (log) posterior mean predictions from both prior (A) and posterior (B) under model 8 to exemplify miscalibrated priors. Notice how the predictions are many orders of magnitude away from one another.

The statistical findings can be summarised as:

- Variable selection was robust to choice of error distribution (Poisson/negative binomial & Gamma/log-normal);

- There is very low residual spatial autocorrelation in the data, leading to unstructured models (OLRE) providing better fit;

- Current state-of-the-art recommended priors for the BYM model lead to a miscalibrated induced distribution on the data $\boldsymbol{Y}$.

4.5 Limitations

In this section I will briefly discuss what I find to be the main limitations of the work presented in this chapter. As already pointed out above, BSSVS has an important limitation in the form of combinatorial multimodality. From a modelling perspective, this makes it difficult to pin down which factors (covariates) are associated with the outcome and obtain reliable estimates of Bayes factors and coefficients. From a computational perspective, multicolinearity-induced multimodality makes the problem intractable and leads to poor exploration of the parameter space by MCMC. As such, the results presented here need to be taken with caution because it is not possible to guarantee that the posterior distribution has been adequately sampled, even if the usual diagnostics failed to detect problems. Additionally, even in the absence of BSSVS ("full" models discussed above), the usual CAR formulation is difficult to sample from in its own right.

Another concern that could be raised it that none of the model comparison metrics employed here seem appropriate to discriminate between models in the present context. Tables 4.5 and 4.7 suggest that the full models, with 57(58) predictors, are the best fitting models. But while one would expect the overall fit to be indeed better for these models, it is expected that the measures such as DIC or WAIC *penalise* bigger models for an execessive number of parameters. Since these big

models are on the verge of non-identifiability, I would argue the results need to be taken with caution.

Finally, I offer a list of improvements that could be made to the current analysis:

- Employ a better, more intuitive parametrisation with better prior calibration, as discussed in Riebler et al. (2016);
- Explore a sparse formulation of the CAR model to facilitate computation as done in Morris (2017);
- Use a continuous covariate (feature) selection procedure along the the lines of Piironen and Vehtari (2017)

I believe these improvements would lead to more stable computation and a better calibrated model which could in turn lead to better epidemiological inferences.

4.6 Conclusions and perspectives

In this chapter I have presented a complete statistical framework to study the association of climatic, socio-economic and genetic predictors with EVD cases and EBOV persistence. I extend the analyses presented in Dudas et al. (2017) for the EVD case data by including the predicted number of viral introductions into each region, and further refine their results by exploiting BSSVS to include a large number of predictors and compute their Bayes factors. Using the posterior predictive distribution, I study which regions that reported no EVD cases would be at higher risk of experiencing epidemics. The modelling of persistence times is the missing piece from Dudas et al. (2017), who looked at case counts alone. I employ the framework presented here to investigate the association of several predictors with viral persistence and also perform predictions in the same fashion as above. I combine the predictions for case counts and viral persistence times to show very

little overlap between areas with high epidemic potential and areas with increased suitability for viral persistence. Since both these factors need to be in place for an outbreak to develop, I argue that these results partially explain why many areas did not report EVD outbreaks even though many of their spatial neighbours experienced widespread epidemics.

Phylodynamic methods allow us to extract epidemiological information from genomic data that would otherwise not be available *via* traditional epidemiological methods. Phylogeography in particular allows for inferences about the latent process of disease spatial spread as infected individuals move in space and then infect others. For the first time we had a rich, densely-sampled, genomic data set that revealed the migration-driven nature of the Ebola epidemic in West Africa, which in practice meant that areas that are neighbours in geographical space need not be strongly connected, whilst areas separated by hundreds of kilometres might be epidemiologically linked. This chapter contains an interesting complementary finding, with implications for disease spatial modelling. None of the spatial models employed here showed better fit to the data, and the fraction of unexplained variation that is spatially structured was small. These results suggest that future disease mapping efforts might greatly benefit from incorporating phylogeographic data in their prior specification. One such way of incorporating migration information is to re-define a neighbourhood structure based not on geographic proximity or sharing of borders but on the viral flow between two regions. Exactly how to incorporate this information and what effect this will have in the fitted models remains an open avenue of future research.

I have left out the issue of phylogenetic uncertainty by assuming that the second order phylogenetic variables (see Table 4.1) were measured without error. In truth, however, these measurements are averages over a distribution of phylogenies and the uncertainty about the phylogenies is not fully accounted for. Ideally, one would want to run the models developed here for many replicates of the variables and evaluate the impact incorporating phylogenetic uncertainty via joint modelling.

Whilst joint modelling of phylogeny and epidemiology remains the ultimate goal, I show that a principled statistical analysis of phylodynamic data (output) can also provide valuable biological insight – see chapter 6 for a more general discussion on the value of separate analysis of phylodynamic data. Traditional epidemiological techniques, when supplemented by phylogenetic/phylodynamic analyses, can lead to a better understanding of the processes shaping pathogen spread within and between populations.

Chapter 5

Investigation of the association between the GP82AV mutation in Ebola virus and fatality rates

> The great tragedy of Science – the slaying of a beautiful hypothesis by an ugly fact.
>
> Thomas Huxley (1825-1895) in *Presidential Address at the British Association, Biogenesis and abiogenesis (1870).*

In this chapter I lay out and expand on my contribution to Diehl et al. (2016). I scrutinise the association between being infected with a particular variant of the virus and the risk of dying, greatly extending the original analyses.

5.1 Introduction

One of the main scientific challenges emerging from the 2013-2016 Ebola virus disease (EVD) epidemic was understanding which factors contributed to such a large-scale epidemic. While many such factors are likely to have been of environmental and socio-economical nature (Dudas et al., 2017), the role of biological adaptation by the virus remains unclear. In particular, evolutionary questions pertaining to the adaptation of the virus to humans and/or accelerated evolution and their impact on the trajectory of the epidemic assumed a central – and often contentious – role in the scientific literature (see e.g. Holmes et al. (2016) and Section 5 in Bausch (2017)).

Ebola virus (EBOV) has a single-stranded negative-sense non-segmented RNA genome of around 18.9 Kb that encodes for at least 8 proteins (genes), amongst which the glycoprotein (GP), nucleoprotein (NP) and polymerase (L) are the most variable and useful for molecular epidemiology. The GP gene codes for the virus glycoprotein, which is expressed on the surface of the viral particle as a transmembrane receptor (Takada et al., 1997). It is thought to be involved with for receptor binding and membrane fusion and thus to be crucial for interaction with the host. This also means the GP gene is one of most studied genes in the Ebola virus (EBOV) genome (Li et al., 2016).

The present chapter concerns the study of the impact of a particular mutation on EVD fatality, namely a non-synonymous C-to-T substitution at nucleotide $6,283$ mutation resulting in the wild-type alanine (A) being replaced by a valine (V) at the 82nd aminoacid position, which interacts with the cell fusion receptor (NPC1). This mutation, henceforth referred to as `GP82AV` emerged early in the epidemic in Guinea (see Figure 1 in Diehl et al. (2016)) and quickly spread through other countries and became dominant whereas the wild type (A) persisted in a few regions. In Diehl et al. (2016), we analysed a set of clinical data for which there was information on which GP82 genotype the virus isolated from a patient had and also the EVD

outcome (died/survived). In that study, a positive association between `GP82AV` and risk of death (odds ratio: 2.09, 95% confidence interval [0.94, 4.64]).

5.1.1 Adaptation to humans

This naturally raised questions of whether the virus had adapted to humans and become more virulent and/or transmissible. Li et al. (2016) argue that no human adaptation occurred during the 2013-2016 Ebola epidemic using phylogenetic methods to show very little variation in evolutionary rates compared to previous epidemics. As the authors themselves note, the evolutionary rate has little value in informing about the likelihood of adaptation to the host (discussed in detail in Holmes et al. (2016)). The authors also estimate selection on the GP gene and find no difference between the West African sequences and samples previous epidemics. Their study did not however include any detailed experimental assessment of mutations in the GP gene and their impact on infectivity[1].

Urbanowicz et al. (2016) and Diehl et al. (2016) – independently – provide detailed experimental evidence on the impact of several point mutations in the GP gene on *in vitro* infectivity in primate and human cells. Both studies employ pseudotyped HIV particles to study viral infectivity of mutants and wild type in several types of cells. These studies showed that viruses containing the `GP82AV` mutation conferred increased human cell entry, and Diehl et al. (2016) found a twofold increase in infectivity. Urbanowicz et al. (2016) also found decreased infectivity in bat cells, suggesting an evolutionary adaptive trade-off. An important observation is that these studies show adaptation to human cells, not to humans[2]. This is an important distinction, because it highlights the fact that even if `GP82AV` confers

[1] In particular, the speculation that "no non-synonymous substitutions occurred on the GP gene coding sequences of EBOV that were likely to affect protein structure or function in any way." (pg. 7) was later challenged by Urbanowicz et al. (2016) and Diehl et al. (2016).

[2] This insightful remark is not my own but was instead made by Vincent Racaniello in his blog (http://www.virology.ws/2016/11/03/increased-infectivity-of-ebola-virus-glycoprotein-from-west-africa/).

higher infectivity and tropism for human cells, it might have very little or no impact on viremia and disease progression.

5.1.2 Considerations about effect size

While the studies by Urbanowicz et al. (2016) and Diehl et al. (2016) provide evidence that `GP82AV` increases infectivity in human cells, the question of whether the increased infectivity could lead to more severe disease remains open. It is entirely possible that the association between `GP82AV` and fatality does not reflect any underlying causation. I expect the outcome, i.e., whether an individual dies of EVD, to be dependent on a swathe of host- and population-specific factors. An individual's age, sex, immune and nutritional status are expected to have a large contribution to risk of death[3]. In addition, whether an individual can get access to medical attention depends on the availability of health care facilities in the region.

It is therefore important to scrutinise these estimates in order to assess their robustness to noise and confounding factors. In other words, the main purpose of this chapter is to present a comprehensive analysis of the data in Diehl et al. (2016) along with an in-depth discussion of the effect size of the `GP82AV` mutation in increasing fatality. The analyses presented in that study are expanded in several ways: (i) I include more covariates in a more sophisticated model; (ii) I assess the robustness of the effect size estimates for `GP82AV` by exploring (a) different model formulations and (b) several prior distributions for the effect size.

I include a discussion about errors in sign (S) and magnitude (M) in the framework of Gelman and Carlin (2014), explore several sources of uncertainty through the principled construction of prior distributions and address heterogeneity between locations using multilevel models. Finally, I employ comparative method techniques

[3]Importantly, this information is unfortunately not available for the data analysed in this chapter.

to estimate heritability of viral loads and disentangle phylogenetic and clinical effects by conditioning on the shared ancestry between the samples.

5.2 Methods

In this chapter I shall make use of both orthodox (frequentist) and Bayesian methods, which might cause some confusion. While I compute both confidence and credibility (or credible) intervals for various quantities, it is important to note that these are fundamentally different objects, in both construction and goal. While confidence intervals are built with frequency guarantees in mind (e.g. a properly constructed 95% confidence interval will contain the true parameter in 95% of replicate experiments, on average), credibility intervals are constructed to give a 95% probability that the true parameter is contained in the interval, without reference to replicated experiments/data sets.

5.2.1 Data

From the 1610 full EBOV genomes available, 316 had information on cycle threshold (Ct) values and 299 sequences had information on patient outcome (died/survived)[4]. The cycle threshold is a measure of how many polymerase chain reaction (PCR) cycles are necessary to obtain amplification of a given target sequence, and thus is directly related to the relative abundance of said sequence in the sample and can be used as a proxy for viral load, or the amount of virus DNA in a sample. Since the cycle threshold (Ct) is inversely and non-linearly related to the amount of virus DNA in the sample, which in turn is thought to be correlated with viral load, I propose to transform Ct in order to get the viral load, `viral_load`$= -\log \mathrm{C}t$. While this transform ensures the sign of the coefficient in a regression setting is easier to

[4]Please note there are small differences between the data described here and that analysed by Diehl et al. (2016).

interpret, magnitude is an entirely separate problem. Since the main goal in this chapter is to study the effect of `GP82AV` I will not pursue this issue any further. There were 233 sequences with complete information on Ct and outcome, of which 221 also had information about location. When focussing solely on sequences from Guinea, we are left with 202 sequences.

In order to aid modelling location-specific heterogeneity in case fatality rates, I collected a number of location-level predictors. Specifically, I collected data on how many health centres, clinics and pharmacies there were in each prefecture (level 2 administrative unit). These data were obtained from the Humanitarian Data Exchange initiative (`https://data.humdata.org/`). In addition, in an effort to account for demographic, economic and epidemiological predictors, I also included information on population size and density, mean gridded economic output and travel time to the nearest settlement with more than $50,000$ inhabitants. To keep coefficients comparable and facilitate prior specification, I transform continuous predictors by subtracting the mean and dividing by two standard deviations (Gelman, 2008). This ensures that coefficients for continuous predictors are comparable to those for binary predictors.

5.2.2 Binary regression

Since the dependent variable is binary, I shall employ generalised linear modelling with a binary outcome $\boldsymbol{Y} \in [0,1]^N$. The model can be written as

$$g(Y_i) = \boldsymbol{X}_i^\intercal \boldsymbol{\beta} + \alpha + \epsilon_i, \tag{5.1}$$

$$\epsilon_i \sim \text{Normal}(0, \sigma^2). \tag{5.2}$$

where $g(\cdot)$ is the *link function* and α is an intercept term. Popular choices of link function include the logistic function $g(p) = \log\left(p(1-p)^{-1}\right)$, the probit link $g(p) = \boldsymbol{\Phi}^{-1}(p)$ – where $\boldsymbol{\Phi}^{-1}(\cdot)$ is standard normal inverse CDF – and the cloglog link

$g(p) = \log(-\log(1-p))$. Each of these has its weaknesses and strengths, ranging from interpretability of the parameters to tail behaviour (Czado and Raftery, 2006). The most common choice by far is the logistic link, mainly because it is easy to interpret parameter estimates; the estimate $\hat{\beta}_i$ represents the marginal log-odds of the $i-$th predictor/covariate. The **odds ratio** $\text{OR}_i = \exp(\hat{\beta}_i)$ gives a direct estimate of risk associated with the $i-$th predictor that is of great value in many scientific fields, particularly Epidemiology (Schmidt and Kohlmann, 2008).

The basic model in (5.1) can be extended in several ways, including particular structures for the errors $\boldsymbol{\epsilon}$. In the following sections I detail multilevel[5] extensions that are useful to the modelling task at hand.

Multi-level models

In several scientific applications, observations are made in batches or groups, e.g. students grouped within schools, disease cases grouped within locations, etc. Multilevel models provide a framework for incorporating information at the individual level (e.g. the age and sex of a measured individual) with data at the group level (e.g. GDP per capita in a city)(Gelman and Hill, 2007). In the modelling situation tackled in this chapter, I am interested in including location-specific covariates that might be associated with risk of death from EVD.

For simplicity assume there is only one grouping factor and that observations come from J groups. Amongst more complicated structures, one can have (group-) varying intercepts, varying slopes or both. For the purposes of this chapter, I will consider

[5]Also called "random effects" models. I prefer the term "multilevel" because it captures the true power of the framework: modelling data at several stages/level and pooling information across strata in a principled way. See sections 1.1 and 11.4 in Gelman and Hill (2007) for a discussion on nomenclature.

location-varying intercepts models, of the form

$$g(Y_{ij}) = \boldsymbol{X}_{ij}^{\intercal}\boldsymbol{\beta} + \alpha_j + \epsilon_{ij}, \tag{5.3}$$

$$\alpha_j = \theta + \delta_j, \tag{5.4}$$

$$\delta_j = \boldsymbol{Z}_j^{\intercal}\boldsymbol{\gamma}. \tag{5.5}$$

where θ is an overall intercept, $\boldsymbol{Z}$ are group (location) level predictors, $\boldsymbol{\gamma}$ are the corresponding coefficients and the errors $\boldsymbol{\epsilon}$ are modelled as before (Eq. 5.1). While this model is said to do *partial pooling*, an alternative model where $\alpha_j \sim f(\cdot)$ assumes locations are i.i.d. and therefore does no pooling of information across locations. A model where $\gamma_j = 0 \, \forall j$ (and therefore $\delta_j = 0 \, \forall j$) is said to be a *complete pooling* model.

Computational details

I fit all generalised multilevel models using the probabilistic programming language Stan (Carpenter et al., 2017), which allows the use of Hamiltonian Monte Carlo (HMC, Neal et al. (2011)) to approximate posterior distributions. I run 4 independent chains of 5000 iterations with 2500 iterations discarded as warm-up. I assessed convergence by visually inspecting parameter traces for stationarity and checking the $\hat{R}$ statistic (Brooks and Gelman, 1998) was close to 1.0.. To ensure appropriate mixing I also calculated effective sample sizes and checked they were above 200.

Prior modelling and effect size

In this chapter I address the impact of prior specification for the regression model parameters in two ways: overall prior calibration, pertaining to the induced distribution of the outcome $\boldsymbol{Y}$ and the impact of several priors for the coefficient of the predictor of interest (absence/presence of `GP82AV`) on the scientific conclusions one might draw from the model(s).

I propose to frame the assessment of effect sizes and prior construction in terms of **risk ratios** (RR), which are easier to understand and interpret. Suppose we are comparing two groups, g_0 and g_1, for the occurrence of a given (binary) outcome. An RR of 1.03 for instance means that individuals in g_1 have a 3% higher chance – *i.e.* risk – of developing the outcome compared to individuals in g_0[6]. Since the models presented above (e.g. 5.1) are parametrised in terms of linear coefficients, we need to transform back and forth between RRs, ORs and coefficients:

$$OR = -\frac{(1-p_0)RR}{(p_0 RR - 1)}, \tag{5.6}$$

$$\beta = \log(OR). \tag{5.7}$$

When modelling risk ratios for fatality, one needs to take into account the baseline **case-fatality rate** (CFR), p_0. The maximum RR achievable for a given CFR is $RR_m = 1 + (1-p_0)/p_0$, and this can be used as an upper bound in the assessment of frequentist estimates and also the construction of priors.

Informative priors

I argue that we can use these functional forms to elicit informative priors for β_{GP82AV} that encode specific hypotheses about the effect of `GP82AV` on fatality rates. The first step in constructing sensible priors is to understand the scale and variation of case-fatality rates. I constructed a Beta distribution for p_0 with parameters $\alpha = 46.063$ and $\beta = 24.511$ using the information presented in page 6 of the meta-analysis by Nyakarahuka et al. (2016). This information suggests the average CFR is $\mathbb{E}[p_0] = 0.65$, which in turn implies that $\mathbb{E}[R_m] = 1.54$, *i.e.* the average maximum risk ratio is about 54%. We can use this information to control the upper (right) tail of the distributions for β_{GP82AV}. I propose we consider three scenarios

[6] Hence $RR = p_1/p_0$, whereas $OR = p_1/(1-p_1) \div p_0/(1-p_0)$.

- **Conservative:** there is very little chance there is an effect. RR should be 1 with credibility interval (CI) $0.95, 1.05$;
- **Neutral:** we are agnostic about the effect; it could be protective or a risk factor. RR should be 1 with 95% CI $0.55, 1.45$;
- **Optimistic:** there is probably an effect of around 20% (RR $= 1.20$), with a 95% CI of $1.10, 1.45$;

To accommodate distributional idiosyncrasies such as different tail behaviour, I construct priors for each scenario using the log-normal and log-logistic families of distributions. I present the results of this elicitation procedure in Figure 5.1.

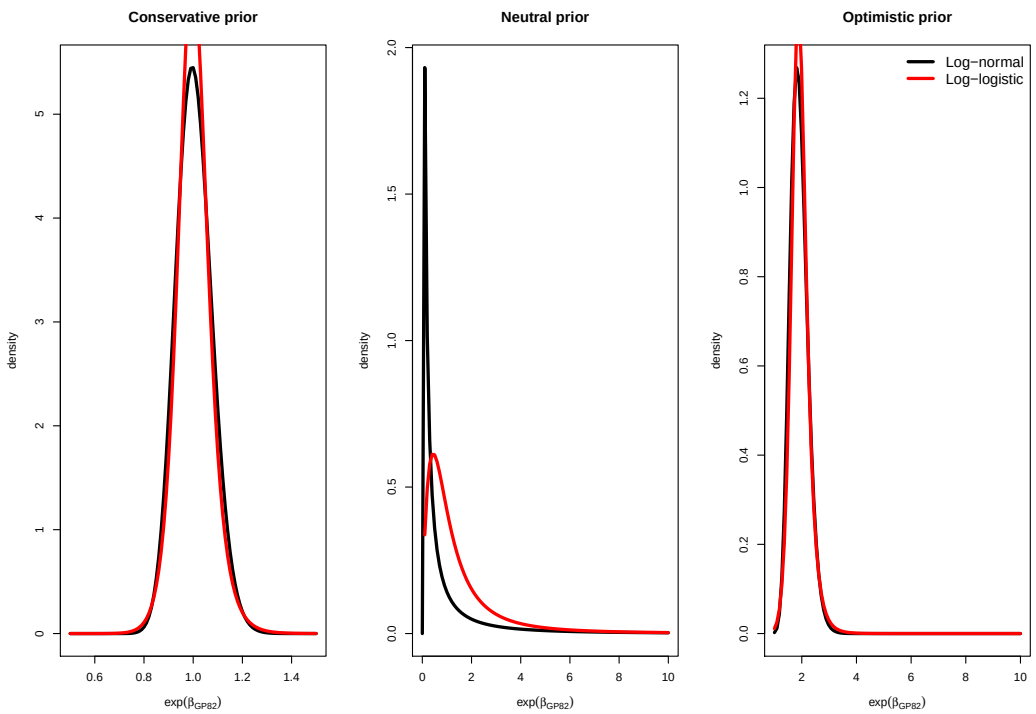


Figure 5.1: Informative priors elicited for three scientific hypotheses about the effect of GP82AV. I show conservative, neutral and optimistic priors constructed using a log-normal (black) and a log-logistic (red) distributions on the odds ratio (OR $= \exp(\beta_{GP82})$) for the GP82AV mutation.

"Default" and weakly-informative priors

When performing inference with informative priors it is also good practice to explore sensitivity to prior specification by employing generic, weakly-informative default priors as a baseline. In this chapter I look at several options for default priors commonly employed in the statistical literature. Firstly, I explore an adaptation of the weakly-informative prior suggested by Seaman III et al. (2012). The authors study an example of logistic regression for coronary heart disease where the continuous predictor is age and they set $\sigma^2 = 25$. Since the range of age is about $100 - 20 = 80$ years, I adapt my priors to have $\sigma^2 = r \times \frac{25}{80} = 1.63$, where $r = 2.59$ is the range of the `viral_load` predictor (after standardisation). I also consider the so-called g-prior (Zellner, 1986), which scales the covariance matrix between coefficients proportional to the variance of the corresponding predictors $\boldsymbol{X}$, as $\pi(\boldsymbol{\beta}) \sim \text{Normal}_P(\mathbf{0}, \frac{1}{g}\boldsymbol{X'X})$. Here I explore a g-prior with $g = N$ and a g-prior with an inverse-Gamma hyperprior on g as discussed in Liang et al. (2008)[7].

A small caveat of eliciting the priors based on three quantities (mean, lower and upper 95% quantiles) is that matching all three is sometimes not attainable. Whenever this happened I chose to prioritise the mean and upper quantile for the neutral prior while focussing on the mean and lower quantile for the optimistic prior. Elicitation of the conservative prior encountered no such problems, mostly due to the small variances involved.

Priors for the multi-level model parameters

I assign a prior on θ such that $q = (1 + \exp(-\theta))^{-1}$ has a Beta distribution with parameters $\alpha = 46.063$ and $\beta = 24.511$. For the no pooling model I assume $\alpha_j \sim \text{Cauchy}(0, 10)$, while for the partial pooling model I assign the location-level coefficients $\boldsymbol{\gamma}$ independent standard Gaussian priors. For the overall intercept θ, I opt for a Gaussian distribution with mean 0 and standard deviation 5. The coefficients ($\boldsymbol{\beta}$ and $\boldsymbol{\delta}$) are given independent Cauchy priors with scale 5/2 by default.

[7]It should be noted that the impact of the g-prior is minimised in our setting due to the scaling of the predictors. It is included here mostly for completeness.

5.2.3 Phylogenetic analyses

Since viral genetic information is available for the clinical data discussed in this chapter, it is of interest to investigate the association between phylogeny – of virus – and the patient's outcome.

For the analyses outlined in this section, I estimated a time-calibrated phylogeny from the 299 complete genomes using BEAST (Drummond et al., 2012). Data were divided into four partitions: coding regions with positions 1, 2 and 3 and intergenic region. I used the HKY model (Hasegawa et al., 1985) model of nucleotide substitution along with Gamma-distributed rate heterogeneity and a log-normal relaxed clock that assumes among-branch variation in rates follows a log-normal distribution.

Continuous trait analysis

The first question I address is whether we can detect heritability in viral loads. I study the association between phylogeny and viral load using a continuous diffusion model of trait evolution, as proposed by Lemey et al. (2010). The idea is to model the trait under analysis as a process the evolves along a phylogeny following Brownian motion (BM), and perform inference by conditioning at the states (values) at the tips. Additionally, I employ the Bayesian approach to Pagel's λ (Pagel, 1999) of Vrancken et al. (2015) to estimate the heritability of viral load. For the purposes of this chapter it is convenient to recall the interpretability of the phylogenetic correlation coefficient λ_B[8]: if $\lambda_B = 0$ there is no correlation between phylogeny and the trait under analysis. Conversely, $\lambda_B = 1$ corresponds to a setting where trait evolution reproduces a Brownian motion process along the phylogeny exactly. Intermediate values represent the lack of adherence of the observed data to a BM. These analyses

[8]The subscript denotes the fact that these estimates are Bayesian estimates derived from the posterior distribution of trees and hence accounting for phylogenetic uncertainty.

used the same phylogenetic model described above, along with a boundary-avoiding Beta prior with parameters $\alpha = \beta = 2$ on λ_B.

Phylogenetic regression

A key assumption of the models presented in Sections 5.2.2 and 5.2.2 is that the observations $\boldsymbol{Y}$ are independent. Since EBOV relies on human to human transmission, that is, all cases are linked through a transmission chain[9], this assumption is clearly violated. One way to partially accommodate this fact is to assume that the phylogeny inferred from the sequences is a reasonable proxy for the dependence structure of the observations. This idea underpins most of the comparative method and is the basis of the phylogenetic logistic regression (PLR) approach of Ives and Garland Jr (2009), which also accommodates binary covariates. This is important because the main focus of this chapter is to study the effect of a binary variable, the absence/presence of the GP82AV mutation.

The idea behind PLR is to use the phylogenetic tree τ to construct a matrix $\boldsymbol{W}$ such that W_{ii} is the distance from tip i to the root and W_{ij} is the length of the branch leading to the last common ancestor of i and j. This matrix is then used to formulate a phylogenetic variance-covariance matrix for $\boldsymbol{Y}$, $\boldsymbol{V}(\alpha)$:

$$\boldsymbol{V}(\alpha) = \boldsymbol{A}^{1/2}\boldsymbol{C}(\alpha)\boldsymbol{A}^{1/2}, \tag{5.8}$$

$$\boldsymbol{C}(\alpha) = \exp\left(-2\alpha\left(\mathbf{1} - \boldsymbol{W}\right)\right), \tag{5.9}$$

where $\mathbf{1}$ is a $1 \times N$ unit vector, $\boldsymbol{A}$ is a diagonal matrix and α is a transition rate parameter controlling the evolution of the process along τ (see Ives and Garland Jr (2009) for more details). I use the `phyloglm()` function in the **phylolm** R package (Ho and Ane, 2014) to fit this model to the data using the phylogeny estimated as

[9]The hypothesis of multiple spillovers from the reservoir has been largely discredited (Baize et al., 2014; Gire et al., 2014).

described above. Confidence intervals were obtained with 500 parametric bootstrap replicates.

5.2.4 Type M and S errors

In the framework of orthodox[10] statistical inference and hypothesis testing, statements regarding reality are framed in terms of contrasts to a hypothesis. The brief exposition of the Neyman-Pearson decision-theoretic paradigm here serves the purpose of motivating the analysis of type M and S errors and in no way reflects the richness of the research in orthodox statistics. Please see e.g. Casella and Berger (2002) and references therein for a more detailed account.

More specifically, one is usually interested in studying the data D under a *null* hypothesis, H_0, which is chosen to encode the sceptical viewpoint. Assuming H_0 holds true, one can use the hypothetical distribution of the data under H_0, $f(d; H_0)$[11], to answer questions such as "how likely would we be to observe values of D or more extreme assuming H_0 were true"? Then, based on a pre-specified threshold for the probability $f(D; H_0)$, one can reject – or fail to reject – the null hypothesis H_0.

These testing procedures are usually calibrated in terms of error probabilities. If one rejects H_0 when it is in fact true, we say a **Type I error** has been committed. The *false positive* rate α is the probability that a testing procedure will lead to a Type I error. In contrast, when one fails to reject H_0 when it is false, one has committed a **Type II error**. The probability that a given testing procedure leads to a Type II error is denoted β. It is useful to define the quantity $1 - \beta$, the **power** of the testing procedure.

[10] Also called frequentist.

[11] Notice I deliberately avoid the conditional probability notation, $f(d|H_0)$. For such a statement to be mathematically correct, one would have assign a valid probability measure over hypotheses, which would in turn lead us strictly outside the boundaries of orthodox Statistics.

Gelman and Carlin (2014) present an alternative perspective to the definition of errors in testing procedures. They propose that we replace the notions of type I and II errors for the more intuitive concepts of errors about the **magnitude (M)** and **sign (S)** of the effect t. The main idea is that if one is trying to estimate the effect t, there are two fundamental mistakes one could make: (i) obtain an estimate with the wrong sign or (ii) obtain an estimate that under- or over-estimates t. Scientifically, a type S error could potentially lead to inferred causal links being reversed, while a type M error could lead to incorrect quantitative statements.

The key quantities in this framework are the measured effect size $\hat{t}$, the measured standard error (s.e.) s and the *hypothesised* true effect size t. From these, one can compute the distribution of the estimate t_r that would be obtained if the data collection and estimation procedure were to be replicated. With the distribution of the replicated estimates t_r in hand, we are then prepared to compute three quantities (Gelman and Carlin, 2014):

- The probability that t_r exceeds the significance threshold, that is, the *power*;
- The probability that t_r has a different sign from t, *i.e.* the type S error;
- The expected type M error, also called the exaggeration ratio, $a_e = E[|\frac{t_r}{t}|]$ which quantifies how far a replicated estimate would be from the true effect size if an statistically significant result were found.

This framework lends itself well to the situation considered here, where we have an estimate of the effect of GP82AV on EVD fatality rates and would like to interrogate the study design and obtain estimates with respect to potential errors of both sign (is the mutation a protective or risk factor or none?) and magnitude (how much more/less risk of death from EVD stems from the absence/presence of GP82AV?).

An important caveat is that I do not know the true effect size t and the literature is scant in terms of information that could be used to establish a reasonable value for

t. To address this I first use the bound derived in Section 5.1 to restrict attention to effect sizes corresponding to risk ratios of less than 1.54. Here I postulate true effect sizes corresponding[12] to risk ratios of 1.05, 1.10, 1.20 and 1.40 and investigate the sensitivity of the calculations.

5.3 Results and discussion

I begin by presenting the contingency table for GP82 genotype (A or V) and EVD outcome (died/survived) in Table 5.1 and then providing raw estimates of the odds and risk ratios and associated standard errors. A raw estimate for the odds ratio can be computed from the number of individuals infected with genotype GP82V that died (a) and survived (b) and likewise for cases of individuals infected with genotype GP82A that died (c) and survived (d). The estimate for the raw odds ratio is then $m_{\mathrm{OR}} = ad/bc$ and a $\alpha\%$ confidence interval can be calculated as $\exp(m_{\mathrm{OR}} \pm \phi^{-1}(\alpha/2) \cdot s_{\mathrm{OR}})$, where $s_{\mathrm{OR}} = \sqrt{(1/a) + (1/b) + (1/c) + (1/d)}$ is the standard error of the estimate. Similarly, one can estimate the risk ratio as $m_{\mathrm{RR}} = c(a+b)/a(c+d)$, with standard error $s_{\mathrm{RR}} = \sqrt{(1/a + 1/c) - (1/(a+b) + 1/(c+d))}$.

These calculations yield a raw odds ratio of 2.57 $(1.34, 4.91)$ with standard error 0.33 and a raw risk ratio of 1.23 $(1.05, 1.45)$ with standard error 0.08 for all of the available data, henceforth called complete data. When restricting attention to Guinea only, we obtain similar estimates of 2.78 $(1.39, 5.54)$ with s.e. 0.35 for the odds ratio and 1.26 $(1.06, 1.49)$ with s.e. 0.09 for the risk ratio.

Table 5.1: Contingency table for GP82AV and EVD fatality – complete data. When restricting attention to Guinea, the numbers become $a = 106$, $b = 18$, $c = 53$ and $d = 25$.

GP82 genotype/Outcome	Died	Survived
V	$a = 130$	$b = 23$
A	$c = 55$	$d = 25$

[12] Please recall that we need equations 5.6 and 5.7 to convert between risk ratios and effect sizes.

It would be desirable, however, to adjust these estimates for the effect of the continuous predictor, `viral_load`. This can be achieved by fitting a logistic generalised linear model with both `viral_load` and `GP82AV` as predictors (see Section 5.2.2), which yields estimates of $\beta_{\texttt{GP82AV}} = 0.75$ $(0.02, 1.49)$ with standard error 0.37, corresponding to an odds ratio of 2.13 $(1.02, 4.44)$[13]. I use these corrected estimates to perform the error analyses in the next section.

From these figures, it would appear that the presence of `GP82AV` is positively associated with fatality, that is, **being infected with a strain of EBOV carrying the V variant increases the risk of dying from EVD**. In the remainder of this chapter I interrogate this claim and attempt to account for multiple sources of uncertainty and confounding.

5.3.1 Experiment 0: type M and S errors for the effect of GP82AV

The first step in scrutinising the claim that `GP82AV` increases mortality from EVD is assessing the probability of various types of error conditional on the measured standard error of the realised estimates.

Table 5.2 shows the potential errors in magnitude (type M) and sign (type S) for the effect of `GP82AV` given the measured standard error (0.37) if the testing threshold for declaring significance were set at $\alpha = 0.05$. Notice how even for reasonably high postulated true effect size (RR = 1.20), the power is low, around 0.40. The probability of inferring an effect of the opposite sign (type S error) is low, and goes down to negligible levels with postulated effects of more than 10% (RR = 1.10). These figures suggest that it is very unlikely that one would estimate the effect of `GP82AV` to be protective, that is, to reduce the risk of dying from EVD. On the other hand, there is substantial chance of overestimation for any true effect below 20% (RR =1.20).

[13]Estimates for the data from Guinea were very similar at OR = 2.35 $(1.10, 5.09)$.

Table 5.2: Magnitude (M) and sign (S) errors for the effect of GP82AV. Assuming different true effect sizes, I show the power, probability of getting an estimate with the wrong sign and the exaggeration factor a_e. I also show the exaggeration factor in the scale of risk ratios, a_e^r, which may be easier to interpret.

Postulated RR	OR[1]	Power[2]	Type S[3]	Exaggeration factor (a_e)	RR exaggeration (a_e^r)[4]
1.05	1.16	0.07	0.143	6.26	1.20
1.1	1.35	0.12	0.025	3.10	1.15
1.2	1.91	0.40	3.63×10^{-3}	1.58	1.07
1.3	2.93	0.81	1.63×10^{-6}	1.12	1.02
1.4	5.44	0.99	3.38×10^{-10}	1.00	1.00

[1]The odds ratio (e^t) obtained by applying Eq. 5.6 with $p_0 = 0.65$.

[2]As defined in Section 5.2.4, assuming an s.e. of 0.37 with 218 degrees of freedom and that the significance level is $\alpha = 0.05$.

[3]Probability that a hypothetical replicated estimate, t_r, has the opposite sign as the true effect size t.

[4]While a_e is the exaggeration in the log-odds scale, a_e^r shows the exaggeration in the risk ratio scale.

The exaggeration in the risk ratio scale is less sensitive to variation in the postulated true effect size, but for a small true effect of 5% the exaggeration would lead to an estimate of $1.2 \times 1.05 = 1.26$ for the risk ratio. Assuming a baseline CFR of 65%, the estimate obtained with a simple logistic regression is $1.23\ (1.00, 1.37)$. While other interpretations are warranted, a conservative perspective would indicate that the realised estimates are compatible with those that would be expected if the true effect size was about 5%. Additionally the uncertainty around the estimates would preclude precise statements, since the confidence intervals suggest increase in mortality due to `GP82AV` could be between 0 and 37%.

5.3.2 Experiment 1: the impact of default priors on a simple logistic regression

Up to this point I have presented results obtained with orthodox (frequentist) methods and from an error-statistical perspective. In this setting, considering the extra flexibility offered by the Bayesian framework can help with incorporating external information and expert knowledge and regularising inference. I now move on to present and discuss Bayesian estimates of the quantities of interest.

As a first step in the Bayesian analysis of the effect of `GP82AV`, I analyse the estimates of β_{GP82AV} obtained using a host of so-called "default" priors. Figure 5.2 shows a prior sensitivity analysis (PSA) for the parameters of a simple logistic regression including only `GP82AV` and `viral_load`. Results show little sensitivity to the default prior used, posterior means and credibility intervals being very similar across all priors considered.

Only the prior constructed according to the recommendations of Seaman III et al. (2012) showed a slight difference in posterior estimates, leading to a 95% posterior credibility interval for the risk ratio of $(1.03, 1.36)$. It is unclear whether the difference in the lower bound (3% against 1%) is of any practical relevance, however. Moreover, these baseline Bayesian estimates are virtually identical to the estimates obtained with ordinary least squares (OLS). This result is not surprising, since the priors employed here are designed to provide regularisation whilst not being too informative about the parameter values. Nevertheless this PSA is useful in establishing a baseline against which all subsequent Bayesian estimates can be compared.

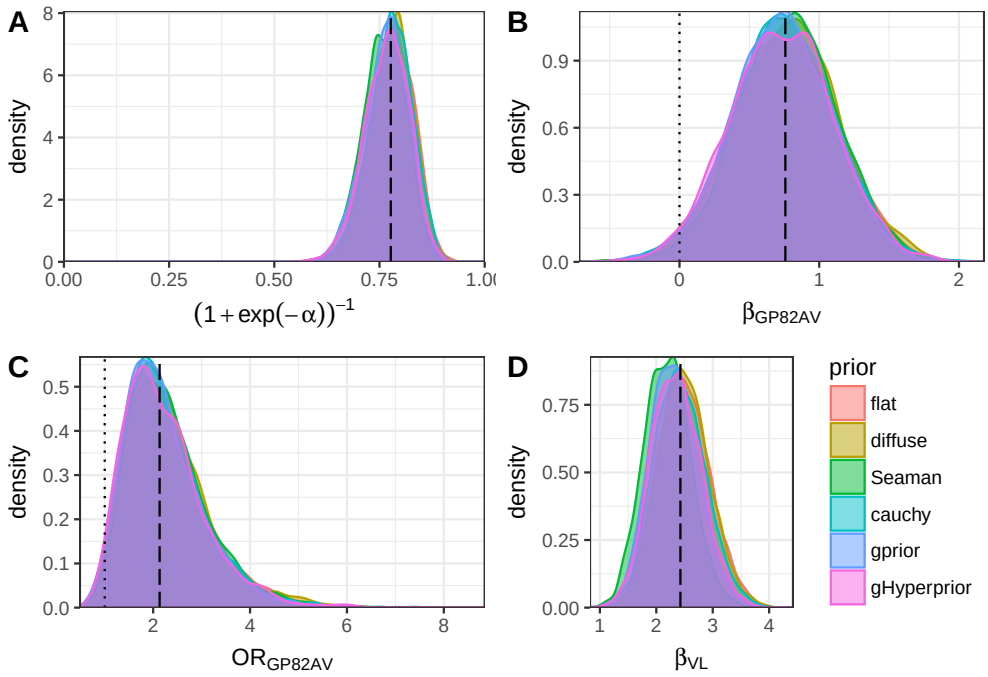


Figure 5.2: Prior sensitivity analysis for a simple logistic regression. In total I considered seven "default" priors routinely employed in the literature (see Section 5.2.2 for details). Panel A shows the posterior distribution for the baseline CFR (a transformation of the intercept α), while panel B shows the posterior distribution for the quantity of greater interest, the effect (coefficient) of GP82AV. I show the resulting distribution on the odds ratio in panel C and the distribution of the coefficient for viral_load in panel D. The vertical dashed line shows the ordinary least squares parameter estimates and the dotted line indicates the null values $\beta_{\text{GP82AV}} = 0$ and $\text{OR} = 1$ for ease of comparison.

5.3.3 Experiment 2: conservative and enthusiastic priors on the effect of GP82AV

One might then be inclined to ask what the effect of incorporating external information and/or expert beliefs into the priors would be. Since external information on the effects of a particular mutation in the virus and disease fatality and expert opinions on the subject are not available, I consider a simplified exercise where I

construct conservative, neutral and optimistic priors for β_{GP82AV}. While Figure 5.1 shows these priors, Figure 5.3 presents the resulting posterior distributions for the parameters of a simple logistic regression.

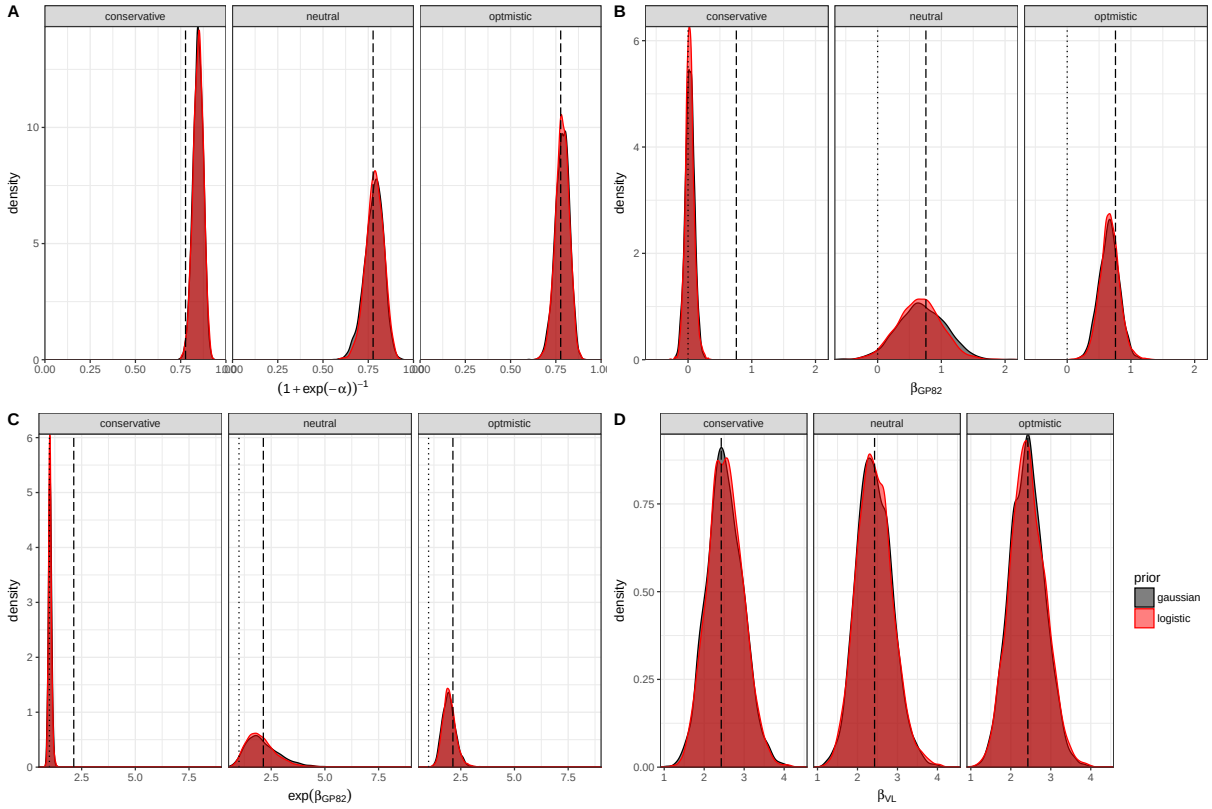


Figure 5.3: Posterior distributions based on informative priors. I show posterior distributions for the parameters of a simple logistic regression under the two families of distributions used to elicit conservative, neutral and optimistic priors for the effect of GP82AV. The posterior for the baseline CFR is shown in panel A, while the coefficient and odds ratios for GP82AV are shown in panel B and C respectively. Panel D shows the posterior distributions for the effect of the viral_load predictor. Dashed vertical lines show the OLS parameter estimates and dotted lines show the null values $\beta_{\text{GP82AV}} = 0$ and $\text{OR} = 1$ for ease of comparison.

Posteriors obtained with different prior families (log-normal and log-logistic) are virtually indistinguishable, indicating that the importance of tail behaviour is negligible in this setting. These sensitivity checks are important in order to gauge

how informative the data are regarding the parameter of interest. In other words, the ability of the data to override (or not) different priors with different tail behaviours is a good indication of the amount of information contained in the data.

The first thing to notice is that the data are unable to change the prior beliefs under the conservative prior: posterior credibility intervals for the RR are virtually identical to the prior credibility interval $(0.95, 1.05)$. As expected, the neutral priors behave similarly to the default priors studied in Section 5.3.2. One key difference, however, is that posterior credibility intervals did include the "null" case of no effect; estimates for the odds ratio were 2.15 $(0.97, 4.10)$ and 2.05 $(0.95, 3.78)$ for the log-normal prior the log-logistic models, respectively. Similarly, estimates for the risk ratio – assuming $p_0 = 0.65$ as before – were 1.20 $(0.99, 1.36)$ and 1.19 $(0.98, 1.35)$. I hypothesise that these results are a consequence of the neutral prior explicitly accommodating the constraints imposed by the CFR, p_0, while the default priors previously considered allow risk ratios much bigger than 1.54.

Somewhat counter-intuitively, in order to construct an optimistic prior – under the two distribution families considered here – one needs to construct priors that have more conservative upper bound. The trade-off is to then have a prior that encodes a positive effect with greater certainty. This is apparent from Figure 5.1 and is reflected in the posterior RR estimates of 1.20 $(1.11, 1.27)$ for the log-normal prior and 1.21 $(1.12, 1.28)$ for the log-logistic.

Experiment 3: mutilevel modelling

When considering the effect of the mutation on fatality rates, the cautious analyst would like to consider alternative explanations connected with the host population rather than the virus. In other words, there might be confounding factors such as differential access to health care across locations, which could be driving the observed association between `GP82AV` and risk of death. I attempt to account for

these factors by formulating a model that explicitly accounts for variation in case fatality rates across locations through the model outlined in Equations 5.3 to 5.5. The analyses in this section pertain to the data from Guinea only, since I could only reliably collect information on population-level predictors for that country.

The results in Figure 5.4 panel C show that there is substantial heterogeneity in case-fatality rates across locations, as evidenced by the bi-modal distribution of CFR for the model with no pooling (red density plot). In panel B we see that assigning each location its own baseline case-fatality rate not only results in larger estimates for β_{GP82AV} but also larger uncertainty, represented by the broader posterior. This is to be expected since there is limited data per location and thus a model with no shrinkage is expected to produce noisier estimates. In addition, it is also likely that the sample data I analyse here is biased towards higher CFR, since these are cases that made it to the health care stage and therefore the sample likely excludes mild cases. When partial pooling is employed, both with and without location-level predictors, we naturally see shrinking towards the overall mean (panel C). The estimated CFR of 0.73 $(0.55, 0.88)$ is still higher than the overall CFR for EVD computed by Nyakarahuka et al. (2016) at 0.65 $(0.54, 0.77)$. According the World Health Organisation (World Health Organization, 2016a), the CFR for the West African EVD outbreak was 0.76 $(0.74, 0.77)$ for Guinea, 0.45 $(0.44, 0.46)$ for Sierra Leone and 0.45 $(0.44, 0.46)$ also for Liberia. Since the sample analysed here (including previous sections) is dominated by Guinean samples, one would also expect to see a higher CFR.

It must be said that while I was able to gain some insight from these analyses, the ultimate goal of including population-level information to help explain differences in probability of death was not achieved. This is because the posterior credibility intervals for the population-level coefficients $\boldsymbol{\gamma}$ included zero for all ten predictors considered, which indicates lack of significant association with the baseline case-fatality rate per location. This finding could very well be the result of a limited sample size compared to the number of groups ($N = 202$, $J = 17$).

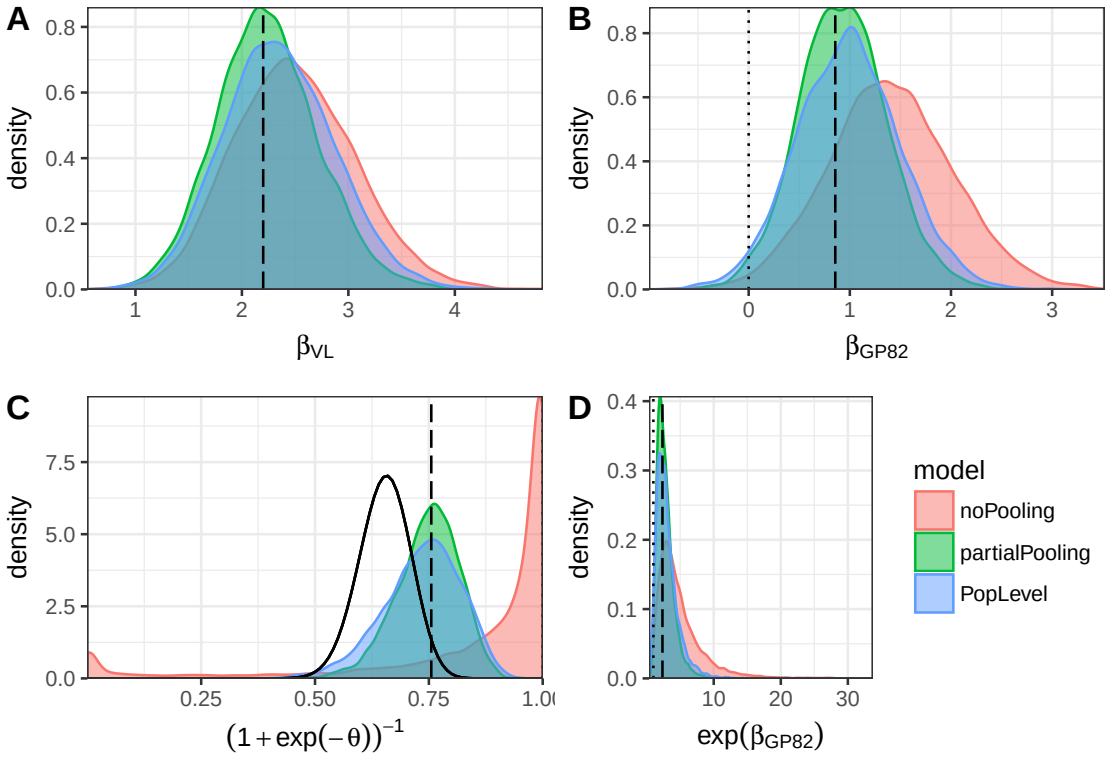


Figure 5.4: Posterior distributions for multilevel logistic models. I compare estimates under three models: no pooling (red), partial pooling without location-level predictors (green) and partial pooling including predictors (blue). Panels A and B show the posterior distributions for the coefficients of viral_load and GP82AV, respectively. In panel C I show the posterior for the case-fatality rates through the appropriate transform of θ. For the no pooling model, I let $\theta = \frac{1}{J}\sum_{j=1}^{J}\alpha_j$. Solid curve depicts the Beta prior I constructed in Section 5.2.2. Dashed vertical lines show the OLS parameter estimates and dotted lines show the null values $\beta_{\text{GP82AV}} = 0$ and OR $= 1$ for ease of comparison.

Experiment 4: accounting for shared ancestry

A key assumption of all the analyses presented so far is that of independence between observations. Since EBOV is transmitted directly, this assumption is fundamentally violated. Despite not being a perfect representation of the transmission history, the phylogeny τ reconstructed from the available sequences is a good proxy for the

dependence structure in the data. Accounting for this dependence is crucial in order to assess the effect of the mutation on fatality rates. When I fitted a phylogenetic logistic regression model to the data using both `GP82AV` and `viral_load` as predictors, I obtained an estimate of 0.76 with 95% confidence interval $(-0.62, 2.13)$ for β_{GP82AV}, 2.13 $(0.54, 8.45)$ for the odds ratio and 1.23 $(0.77, 1.45)$ for the risk ratio. It is clear from these estimates that once we account for the dependence between observations (sequences), the uncertainty about the estimates definitely precludes strong statements about the effect of `GP82AV`.

The analysis of the association between `viral_load` and phylogeny using the framework of Vrancken et al. (2015) yielded $\lambda_B = 0.53$ $(0.29, 0.79)$. These estimates point towards a moderate degree of phylogenetic signal, *i.e.* phylogeny-trait correlation. In addition, the PLR analysis described previously also yielded estimates of the Ornstein-Uhlenbeck variance-restraining parameter α which are consistent with a significant association between the trait and phylogeny – mean: 5.63; 95% parametric bootstrap interval: $(1.32, 76.33)$ – despite a considerable amount of uncertainty.

An important observation is that from all the models considered in this study, higher viral loads are unequivocally associated with higher fatality rates. Figure 5.5 shows clear structure in viral load values across lineages, in agreement with the previous findings. Taken together with the results from this section, this provides a possible explanation for the observed results: since viral load is moderately inheritable and the V mutation in GP emerged quite early – in a deep branch –, the apparent association between `GP82AV` and fatality might just be a result of latent dependence on the tree. These findings reinforce the claim by Russell and de Jong (2017) that the study and management of infectious diseases must be evolutionary, that is take into account evolutionary factors, the main of which is shared ancestry.

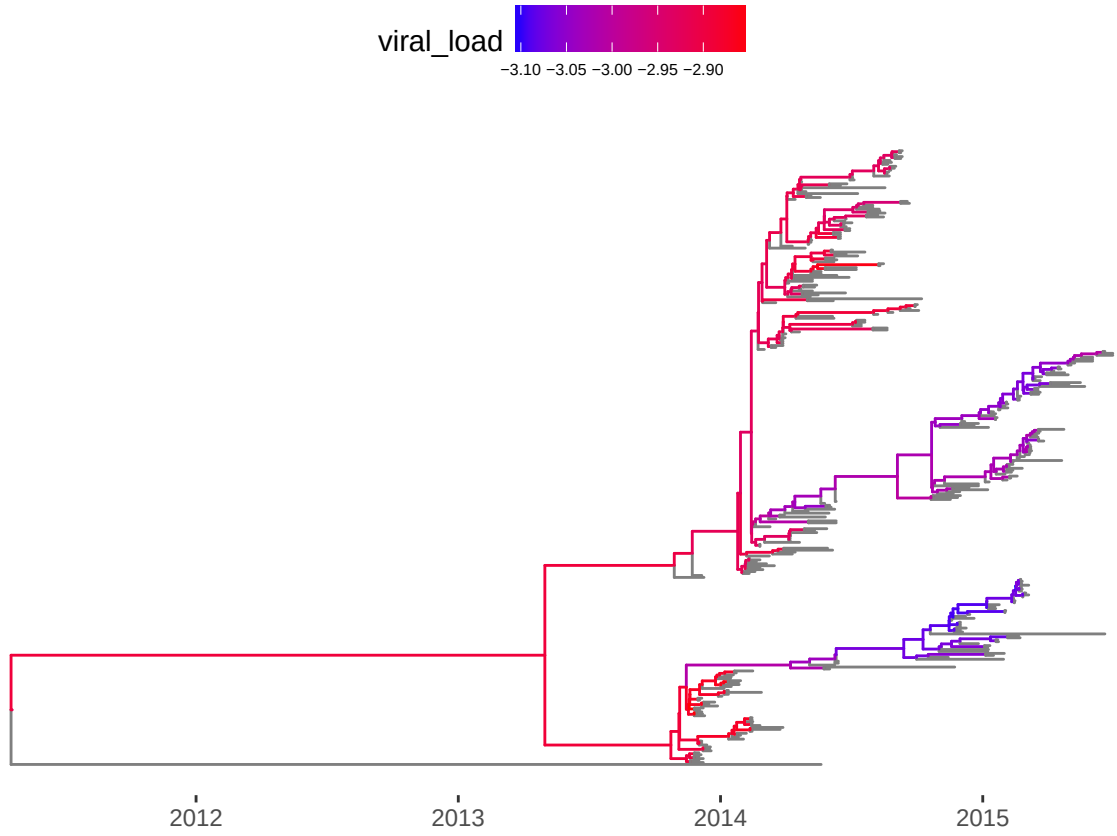


Figure 5.5: Time-calibrated phylogeny annotated with viral loads. I show inferred (posterior median) viral load values at internal branches using the Brownian motion continuous trait evolution model outlined in Section 5.2.3. Gray edges correspond to external nodes (tips).

5.3.4 Does GP82AV increase the risk of death from EVD?

The short answer is: probably not.

The first issue to consider is that of **uncertainty**. In order to make claims about the effect of the `GP82AV` mutation one needs first to quantify the uncertainty about the effect size. As the analyses in Section 5.2.4 show, for a low true effect size (risk ratios < 1.10), the standard error estimated with the current design/data could lead to gross overestimation of the effect. It must also be said that these analyses

suggest that a sign error, that is an error that would lead to inferring `GP82AV` to be a protective factor, is very unlikely. If the mutation has any bearing on the outcome of EVD at all, it most likely increases the risk of dying.

With regard to the uncertainty about the baseline case-fatality rate (p_0), panel A in Figure 5.6 shows the uncertainty about the maximum effect of the V mutation compared to the wild-type. From this it is clear that it is very unlikely that being infected with a virus carrying the V mutation increases the probability of dying by more than 60%. Panel B shows that the original estimate for β_{GP82AV} lies in a region of large uncertainty w.r.t. risk ratios. This means that, considering the variation (uncertainty) of the CFR, any estimate of the risk ratio corresponding to the region of the observed log-odds ratio (taken at face value) could be between 1.15 and 1.35 with probability 95%.

A second point to consider is that of **heterogeneity**. From the results in Section 5.3.3 we see that there is considerable heterogeneity in fatality rates across locations, despite none of the included location-level predictors having a strong association with the baseline CFR. When we account for heterogeneity across locations, we see that the effect of `GP82AV` gets overshadowed by the uncertainty induced by the heterogeneity (Figure 5.7, left panel). While there is still substantial separation between mean responses for both variants (GP82A and GP82V), the credibility intervals overlap considerably, preventing strong claims.

The third point I propose considering is **dependence**. For instance, the methods in Section 5.2.2 rely heavily on the translation of risk ratios into odds ratios in order to elicit the prior distributions. But as Morozova et al. (2018) argue, risk ratios need careful interpretation when the outcome is contagious because transmissibility induces dependence between individuals. The prior elicitation conducted here ignores this problem, although this could be relaxed in future research. While exploring different sources of error, prior formulations and accounting for heterogeneity are

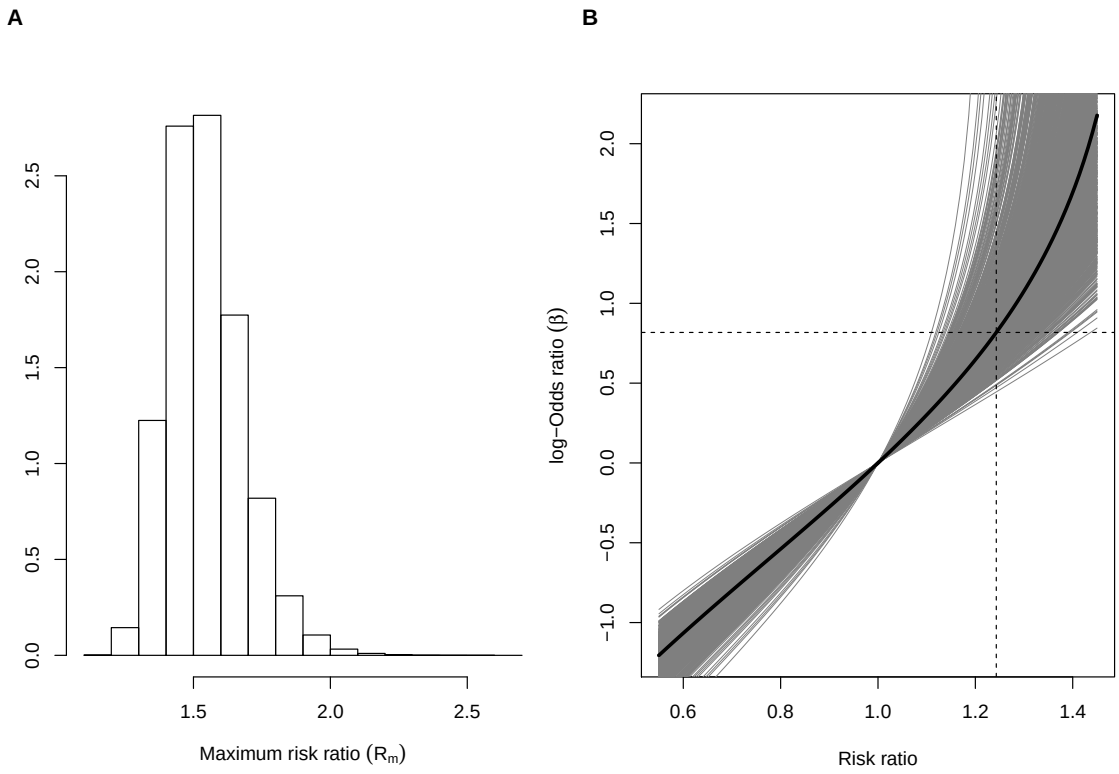


Figure 5.6: Induced distributions on quantities of interest from the uncertainty about the case fatality ratio (CFR), p_0. Using the distribution for p_0 constructed using the information in Nyakarahuka et al. (2016) I explore the induced distributions on quantities of interest. In panel A I show the distribution of the maximum risk ratio (R_m). panel B shows the relationship between (log) odds ratios, *i.e* the coefficient β, and risk ratios (eqs. 5.6 and 5.7) through 1000 replicates from the distribution of p_0 (light grey lines). Black solid line shows the relationship for $E[p_0] = 0.65$ and the dashed lines show the estimated coefficient for GP82AV ($\hat{\beta}_{GP82AV}$) and the corresponding risk ratio – again with $E[p_0] = 0.65$.

all important, accommodating dependence between observations addresses a key assumption of my previous analyses. Having access to complete genomes instead of, say, just genotyping information, allows us to estimate a phylogeny for the data points in study and hence obtain a proxy for the underlying dependence structure of the data. The results from Section 5.3.3 show that, after controlling for shared

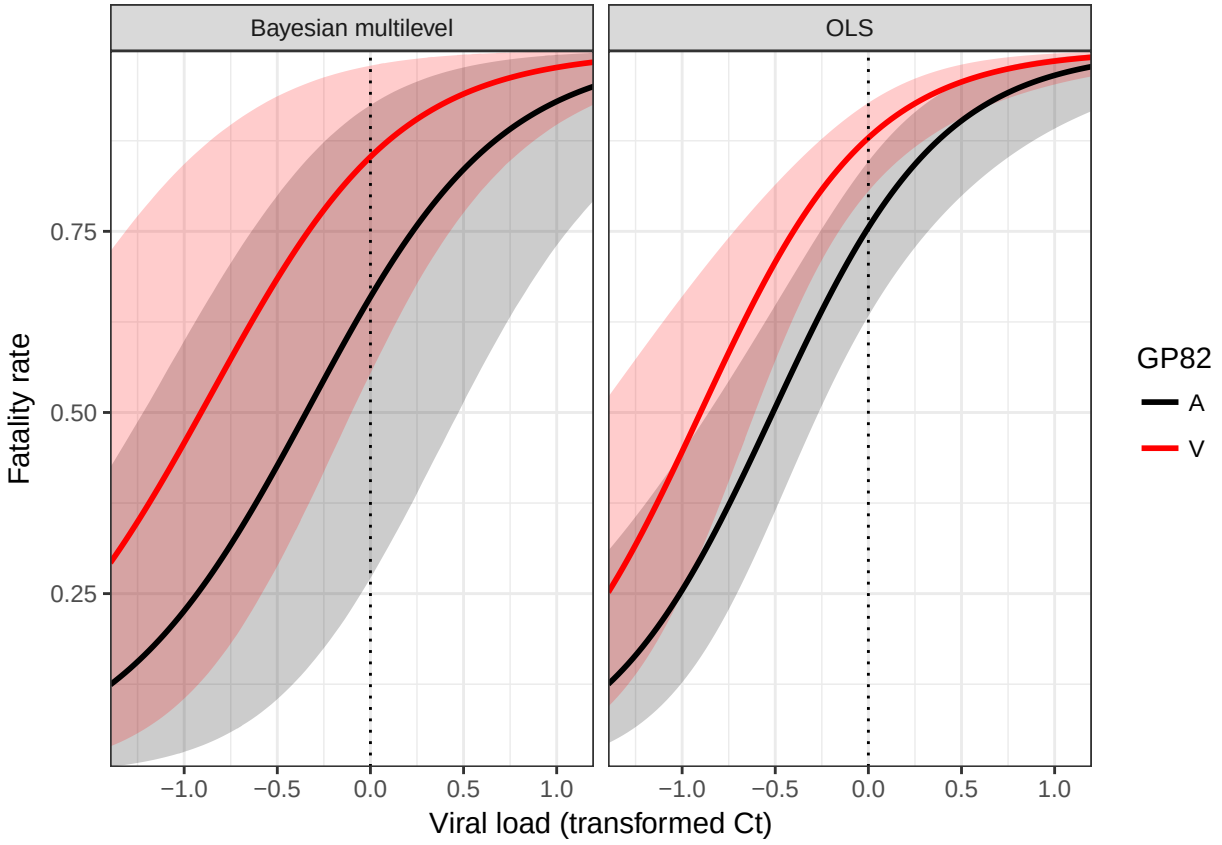


Figure 5.7: Predicted case fatality ratio curves from OLS and Bayesian multilevel models. I show the predicted case-fatality rate curves for viral load (transformed Ct) for both genotypes (A: black, V: red). Vertical dotted line indicates the average viral load (recall the predictor is standardised) for ease of comparison.

ancestry, the (bootstrap) confidence intervals for the effect of `GP82AV` include the null hypothesis of no effect.

While I could combine all previous approaches into one single complicated model for further inquiry, it is my opinion that what we really need is more data and/or a better design. Two final points worth considering are the roles of host heterogeneity and study design. The data I analyse in this chapter unfortunately do not include important host (patient) information such as age, sex, pre-existing health conditions, etc. It is entirely conceivable that heterogeneity in immune responses could be

driving most of the variation in fatality rates. Future efforts in matching patient data and available sequences may alleviate this problem by providing the missing information.

Even if more detailed patient information were available, however, the sample analysed here is one of convenience, collected as the epidemic progressed and patients were treated and tested at health care facilities, without a careful experimental design in mind. This introduces biases that are hard to correct for (see for example the results in Section 5.3.3). A key insight is that if a mutation appears early on during an epidemic and then becomes fixed, it is difficult to disentangle its effects from other confounding factors at various levels, (exposure, geography, control measures, etc, see discussion in Diehl et al. (2016)). Conversely, if a mutation occurs repeatedly during an epidemic, we have much more scientific power to study its association with outcomes of interest such as fatality. In other words, whether `GP82AV` does indeed increase the chance of dying of EVD may not be answerable insofar as it may require direct experimentation in human subjects, which would be ethically unacceptable. A recent study detailing experimentation in non-human subjects (mice and rhesus macaques) failed to show increased pathogenicity of `GP82AV` (Marzi et al., 2018).

5.4 Conclusions

In this chapter I lay out a principled analysis of the effects of a viral mutation on the fatality of the host. I tackle the issue from multiple angles and, crucially, consider several sources of uncertainty. When many sources of uncertainty about the effect of the `GP82AV` mutation are considered, it is clear that strong claims are not warranted. I also show that accounting for dependence between observations through the phylogeny leads to a substantial increase in uncertainty about the effect.

Chapter 6

Discussion

> Nobody exists on purpose,
> nobody belongs anywhere,
> everybody is going to die.
> Come watch TV.
>
> Morty Smith (Dimension C-137)
> to his sister Summer, in episode
> "Rixty Minutes" of *Ricky and*
> *Morty* (2014).

In this final chapter I offer critical appraisal of the work I presented in this thesis and discuss how my findings fit into the general picture.

6.1 Improving the methodological apparatus

A framework for the development and testing of proposal mechanisms

The fist goal of this work was to expand the set of tools available to the practitioner when running MCMC in phylogenetic space in order to reconstruct phylogenies.

When employing complex phylodynamic models to extract useful information from genomic data, phylogenies are frequently a *nuisance parameter*, that is a parameter that needs to be estimated but is not of direct (scientific) interest. Therefore when performing posterior simulation through MCMC, special focus is given to diagnosing convergence with respect to continuous parameters such as the evolutionary rate, growth rates and the basic reproductive number, R_0. This is a crucial insight, only partially explored in this thesis: while we would like to obtain the most accurate representation of the posterior distribution of trees, ultimately we are concerned with designing algorithms that can efficiently traverse phylogenetic space so that we can integrate over plausible phylogenies and hence obtain estimates that accommodate phylogenetic uncertainty. In other words, since we are interested in, say, the evolutionary rate, how much (computational) effort is necessary to obtain marginal distributions for this parameter that properly accommodate phylogenetic uncertainty? This is mainly achieved through the design of efficient transition kernels that lead to rapidly mixing Markov chains.

As argued in Chapter 2, simple, unguided random proposal mechanisms for phylogenies usually have low acceptance probabilities and show poor performance. On the other hand, any "guided" candidate-generating mechanism (operator) needs to be cheap enough to compute so as to enable fitting models in reasonable time, a task made even more pressing in an era when researchers are moving toward real-time analysis of phylodynamic data. Our solution entails an adaptive operator, SubTreeLeap (`STL`) that simultaneously updates branch lengths and tree topology while exploiting a natural metric in phylogenetic space, patristic distance. I show that this operator leads to better mixing not only in phylogenetic space but also for continuous parameters that depend on the phylogeny. Results were particularly encouraging for a very challenging data set of 1610 complete EBOV genomes. I also found, however, that `STL` can suffer with premature adaptation, leading to the chain being stuck at a mode and hence poor sampling of the target. This issue is almost completely ameliorated by combining `STL` with `STJ`, another operator developed in this thesis and that can help the chain jump between modes.

Chapter 2 also details some new tools I developed in order to better evaluate MCMC performance for phylogenies. For example, I expand on the the idea put forth by Lanfear et al. (2016) and propose obtaining a golden tree from independent, long golden runs (Höhna et al., 2008; Lakner et al., 2008) and then computing the distribution of the distance to this golden tree under several metrics and then using these distributions as a (univariate) representation of the desired target. This allows the analysis of warm-up (burn-in) times and also mixing performance. Being inherently based a global metric, however, assessing convergence based on the distance to true tree is likely to be stringent, similar to comparing global convergence in terms of overall clade frequencies. Whilst previous studies (e.g. Lakner et al. (2008) and Höhna et al. (2008)) proposed looking solely at clade frequencies, I propose to augment the analysis of clade space by considering the rate at which clade indicators flip in the chain. While not without its technical difficulties – for instance it depends on knowing the true clade frequencies, which is frequently infeasible in practice – this quantity offers yet another tool in the analyst's toolbox for determining whether the MCMC is reliable. Section 2.2.3 in Chapter 2 offers some tests and checks one can implement to verify the correctness of a phylogenetic sampler in the absence of an independent implementation of the sampler in question but when a golden standard implementation is available. These tools are important because phylogenetic space is a non-standard parameter space, which renders most common tools for assessing convergence and correctness partially or completely inappropriate. I hope these tools can be combined with the existing tool set and used in future Bayesian phylogenetics studies.

While the improved performance shown by `STX` (`STJ` + `STL`) is encouraging, specially considering the results for the EBOV data set, I see the efforts in Chapter 2 as a starting point rather than a complete solution. There are further performance gains to be accrued by more carefully considering which aspects of the structure of the target distribution we can exploit (see below and discussion in Chapter 2 for more). For instance, `STL` finds destinations a certain distance δ away from the current node but then picks amongst them uniformly. Would picking locations proportional to

their distance lead to better performance? Or would it exacerbate the premature adaptation problems already observed? While theoretical investigation is certainly a possibility to be explored in the future, these questions ultimately need to be settled by (computational) experiment.

We need a better mathematical understanding of phylogenetic space

As I have shown in Chapters 2 and 3, representing phylogenetic space through multi-dimensional scaling can lead to inconsistent results. For instance, MDS under the Steel-Penny metric (Steel and Penny, 1993) failed to show serious problems with the EBOV 1610 runs using only `STL` (Chapter 3, Figure 3.5). In addition, while MDS analyses routinely show only the first two components of the projection (presumably for ease of interpretation/visualisation) for some of the data sets I analysed in Chapter 3 – and some metrics – it is clear that one needs more than two components to fully capture a substantial amount of the variation in the data. While MDS remains a useful tool for studying exploration of phylogenetic space, these results should heed a warning to practitioners that (i) multiple metrics should be explored and (ii) bi-dimensional representations might not be sufficient to adequately represent the space under study.

On the theoretical front, compared to what is known for high dimensional smooth parameter spaces, mainly manifolds, we know very little about the geometry and properties of phylogenetic space. The efforts by Billera et al. (2001), Gavryushkin et al. (2018) and Whidden and Matsen (2017) at characterising the space of phylogenies, while extremely valuable, only scratch the surface. Phylogenetic space admits both discrete ("discrete tree space in St. John (2017)) and continuous ("continuous tree space) representations, none of them canonical. While these capture distinct features the inter-dependence between topology and branch lengths means that any geometrical understanding of phylogenetic space needs to come accompanied by some natural traversing metric. The Billera-Vogtmann-Holmes (BHV) (Billera et al., 2001) space admits unique geodesics and allows for the

definition of quantities such as expectations and variances, and appears as a strong candidate for a canonical representation of phylogenetic space. The "stickiness" of the mean – that is perturbing a sample of trees may very well not change their mean tree – and the sharpness (lack of differentiability) at the boundaries are important shortcomings, however.

Very recent developments have brought the promise of a more a natural representation of phylogenetic space in $\mathbb{R}^d$ in the form of the log map of Barden et al. (2018). The log map allows for a representation of the BHV space that not only allows for expectations and variances but also admits rotations[1] and thus opens up the possibility for computing (defining) covariances and hence prove central limit theorems in this space. This new representation has already been used successfully to quantify uncertainty in phylogenetic estimates in a paper by Willis and Bell (2017). It is possible that these new results could be combined with the work of Nye (2015) on Brownian motion in phylogenetic space to design more efficient sampling methods. The crucial insight is that, as argued by St. John (2017), a better understanding of the structure of phylogenetic space can not only improve search algorithms – be them for optimisation or sampling – but also the analysis of MCMC output. The core idea here is that we need better **statistical** representations (see Holland (2013)) of phylogenetic space so we can leverage tools from data analysis and visualisation to study phylogenies.

6.2 Data integration in phylodynamics

Phylodynamics is ultimately concerned with using genomic information to answer epidemiological and evolutionary questions. This means that it is equally as important or more to develop ways of integrating data from several sources in order to draw inference about the processes driving pathogen evolution and spread. As

[1] And preserves directionality.

the quantity and quality of data grow, so too must our statistical and mathematical models, if we are to have any hope to capture nature's intricacies in any useful way. I argue that data integration can happen in mainly two forms: (a) via the construction of **joint models** that explicitly incorporate the relationship between phylogeny and other model parameters and (b) via the analysis and modelling of the **output of phylodynamic analyses**. Both applications in this thesis are of the latter type.

In Chapter 4 I explored a fairly standard generalised linear model coupled with Bayesian stochastic variable selection (BSSVS) to investigate factors associated with the proliferation (number of cases) and persistence of EBOV in West Africa. There are two points at which the output of a phylodynamic analysis enters this model, first in the form of phylogenetic covariates such as the number of viral introductions and distances between introductions which were extracted from a phylogeographic model fitted to the EBOV data (Dudas et al., 2017). The second way phylodynamic data is combined with other environmental and socio-economic data in Chapter 4 is by using persistence times (again extracted from the data of Dudas et al. (2017)) as a response variable. Using this information allowed me to discover that the factors associated with persistence are different than those associated with the number of cases, which is valuable epidemiological inference that could be used by public health authorities. It should be noted that the application of the BSSVS framework bears many similarities with the model employed to investigate which factors contributed to the *spread* of EBOV which is a joint model in the form (a) above. In this particular application while it would have been quite difficult to construct a full joint model in BEAST, accommodating phylogenetic uncertainty is straightforward and entails running the model for samples extracted from each iteration of BEAST run. This constitutes a possibility for future research.

The main concern in Chapter 5 was to properly accommodate uncertainty – not just phylogenetic – in order to evaluate the association between a particular mutation on the virus and the outcome of EVD (Diehl et al., 2016). This was accomplished

by tackling the same question from several angles, employing different statistical models. I used a phylogenetic Brownian motion model to study the relationship between viral load and phylogeny, as a way of gaining insight into how this covariate – which is strongly associated with the outcome – varies along the underlying phylogeny. In addition, I fit phylogenetic regression models that account for shared ancestry between observations and therefore are conceptually superior to i.i.d models in this particular setting. In this instance, I could have built a full joint model in a manner similar to the approach of Cybis et al. (2015), who build a latent liability model that models a latent variable through Brownian motion and apply the appropriate transformation in order to be able to condition on the data at the tips[2].

As stated in Chapter 1, proper appreciation of phylogenetic uncertainty, more specifically propagating phylogenetic uncertainty to parameter estimates, is at the heart of Bayesian phylodynamic inference. On the other hand, joint modelling of phylogeny and the process of interest (transmission, growth, virulence, adaptation, etc) is sometimes hard to implement. This suggests that, while not ideal, separate statistical treatment of phylodynamic output as exemplified in Chapters 4 and 5 may offer a helpful compromise between full joint modelling and not acknowledging phylogeny altogether. As discussed by Baele et al. (2016), even when we know for a fact that there is shared ancestry between observations, the researcher should ask herself whether the variables under study are correlated with phylogeny[3], what can be partially – if sub-optimally – addressed by estimating phylogenetic signal (Vrancken et al., 2015) and quantify the correlation between covariates and phylogeny. This questions is – perhaps unsurprisingly – connected to the question posed above about the strength of dependence of certain parameters of interest $\boldsymbol{\theta}$ on the underlying phylogeny: if the joint model results in virtually

[2]In this case any of the link functions commonly associated with binary regression would be suitable.

[3]Recall that it is entirely possible to have non-independent but uncorrelated random variables.

independent components between phylogenies and $\boldsymbol{\theta}$, then there is very little to gain by constructing a joint model, and the technical overhead is not negligible.

I should stress however that, in principle, a full joint model is always preferable. Baele et al. (2016) not only offer an excellent review of the state of the art in phylodynamics data integration but also show a plethora of examples for which the technical machinery is already in place. These range from compartmental epidemiological models (Rasmussen et al., 2011) to transmission trees (Hall, 2015) to antigenic evolution (Bedford et al., 2015). Their review touches on many points not covered in this section.

6.3 Where are we headed?

I conclude this chapter with a few predictions on where phylodynamics might go in the coming years, focusing mainly on questions related to those tackled in this thesis.

Metropolis-Hastings is not the ultimate tool

Developed in the 1950s and popularised in Statistics after the 1970s, the Metropolis-Hastings (MH) algorithm has been a work horse of computational statistics ever since, much due to its simplicity and ease of implementation. Phylogenetic space presents a challenge to traditional MCMC because of its mixed geometry, including discrete and continuous components that interact in non-trivial ways rendering most of the available theoretical results inadequate. Perhaps of because of this non-standard nature of the parameter space, MCMC for phylogenetics has been mainly restricted to Metropolis-Hastings schemes. BEAST (Drummond et al., 2012), BEAST2 (Bouckaert et al., 2014), Mr Bayes (Huelsenbeck and Ronquist, 2001) and RevBayes (Höhna et al., 2016) all rely heavily on MH-type updates to sample from the posterior.

While MH allows sampling from complex distributions with relative little implementation overhead, it has become clear over the years that it is also prone to random walk behaviour, specially in high-dimensions. This is because of a phenomenon known as *concentration of measure*, which loosely means that as the dimension grows, less and less mass is concentrated around the mode(s). Therefore, any transition kernel (proposal mechanism) that does not exploit the structure of the target but instead naively proposes points in the high-dimensional space based on perturbations of the current state is going to mix poorly. Since most update schemes implemented in the above-mentioned packages are based on random perturbations of the current phylogeny, I argue that there is much room for improvement within the MH framework. As discussed more thoroughly in the discussion section of chapter 2, while our proposed kernels, `SubTreeJump` and `SubTreeLeap`, enjoy many convenient theoretical and empirical properties, they ultimately fail to exploit posterior structure to its fullest. Since phylogenetic space is inherently multi-modal, the development of mode-jumping (Tjelmeland and Hegstad, 2001) and locally-balanced (Zanella, 2017) transition kernels is bound to lead to substantial gains.

However, whilst new transition kernels for MH are likely to continue yielding substantial performance gains, the algorithm itself can only go so far. I therefore argue that in parallel to the development and testing of new MH-based MCMC schemes, we should also considering alternative sampling algorithms, such as Hamiltonian Monte Carlo (HMC). HMC has become a very popular MCMC algorithm as of late, mainly due to its superior performance for complex models (multi-level, Gaussian processes, etc) that were previously hard to fit via Gibbs samplers and MH. The chief limitation of HMC for sampling phylogenetic posteriors is that it relies on constructing a Hamiltonian for which derivatives w.r.t. the parameters exist and are continuous. A recent paper by Dinh et al. (2017) gets around this limitation by constructing a so-called "Probabilistic Path Hamiltonian Monte Carlo" (PPHMC) algorithm that exploits the continuous nature of orthants in BHV space and carefully constructs transitions when traversing the boundaries of the space, which correspond to different topologies. While their results are still

preliminary and a substantial amount of computational work would be necessary to approximate the types of posteriors that BEAST can sample, their work does bring the promise of leveraging several recent theoretical developments made for HMC to sample phylogenies.

The main insight missing at the moment is what constitutes a good transition in phylogenetic space. Without this understanding, researchers often resort to uninformed random perturbations of the current phylogeny, as is the case for boundary (topological) moves in the PPHMC of Dinh et al. (2017). We need more empirical study of real posteriors coupled with theoretical investigations in order to design transition kernels that are effective.

Towards a robust Bayesian phylogenetic workflow

From a statistical point of view, it is desirable that any inference procedure and its associated computational machinery have checks for self-consistency that allow the user/practitioner to detect when results are valid. Ideally, when a researcher inputs her data into a program such as BEAST with the goal of performing a phylodynamic analysis, she should be able to detect problems with her model and/or data while running the program instead of receiving seemingly valid answers that are ultimately wrong. A good example is model selection via marginal likelihood estimation: if a user attempts to estimate marginal likelihoods using improper priors, the program should at the very least throw a warning, since there is no guarantee the posterior will be proper and hence that the results will be valid. As the field moves forward and more researchers adopt the Bayesian framework, I envisage the development of a robust Bayesian phylogenetics workflow as a crucial tool to ensure valid scientific inference.

Running multiple chains starting from overdispersed states is a powerful tool to detect glaring problems, but to the best of my knowledge, this is not common practice amongst biologists using BEAST. In contrast, Mr Bayes runs multiple chains by default and computes the average standard deviation in clade (split)

frequencies (ASDSF) between chains. Ideally, any Bayesian phylogenetics workflow would not only include multiple runs, but also employ more diagnostic metrics beyond ASDSF. Computing the potential scale reduction factor for phylogenies (under different metrics) and continuous parameters is straightforward and should be included by default in programs such as Tracer (`http://tree.bio.ed.ac.uk/software/tracer/`) for the latter.

MCMC might present subtle biases that are hard to diagnose (see Chapter 3). These bias might stem from multi-modality and extreme (posterior) correlations that make it hard for the MCMC to explore parameter space efficiently, but not always lead to detectable problems such as low ESS. I therefore argue that more theoretical statistical work needs to be done in order to understand phylogenetic posteriors, specially regarding prior modelling. The work by Yang and Rannala (2005) and Wang and Yang (2014) has called attention to the construction of priors in phylogenetics, where it is not uncommon for researchers to assign uniform, seemingly "uninformative" priors that lead to very strong constraints on the posterior and have unintended consequences (Seaman III et al., 2012). The coalescent has theoretical justifications that make it an attractive prior on phylogenies, but as I show in Chapter 2 (Section 2.2.2), the prior is flat in (potentially large) portions of the space of ranked phylogenies[4], which means it places no constraints on the potential multimodality induced by the likelihood. Moreover, Boskova et al. (2018) have very recently shown that phylodynamic analyses of early epidemics can be quite sensitive to tree priors. It is clear, therefore, that the development of better priors, that incorporate more structure and are less susceptible to biases, is an important avenue of future research (see Möller et al. (2018) for a step in this direction).

Real-time phylodynamics

On the applied front, recent advances in sequencing technology allowed researchers to obtain genomic sequences with minimal delay (Quick et al., 2016). Coupled with

[4]I stress however I do not claim originality about this observation.

successes in the analyses of large collections of serially-sampled viral genomes (Dudas et al., 2017), this has opened the possibility of real-time analysis of pathogen genomic data as the epidemic progresses.

Initiatives such as `nextstrain` (`http://nextstrain.org/`, Hadfield et al. (2017)) have shown that it is possible to not only curate existing information publicly available on GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) but also provide powerful visualisation capabilities that combine the output of complex phylogenetic/phylodynamic models running under the hood with maps and colours to display information in usable form. Recently, the `ARTIC` network (`http://artic.network/`) has been formed to develop, in their own words, "an end-to-end system for processing samples from viral outbreaks to generate real-time epidemiological information that is interpretable and actionable by public health bodies."

Apart from further development in the logistics of deploying the sequencing apparatus in the field – the "lab in a suitcase" – the success of this initiative will hinge on having the computational capabilities for processing the data (Bioinformatics), drawing inference (Statistics) and conveying information in a useful way (visualisation). On the statistical side, sequential Monte Carlo (SMC) methods are at the forefront of Bayesian on-line phylogenetic inference (Dinh et al., 2016; Fourment et al., 2017; Everitt et al., 2018). Much conceptual work is needed in order to ensure that phylogenies built when one (or a few) taxa are added at a time are accurate and I see the development of heuristics for taxa placement as a promising way of shortening convergence times in sequential analyses.

These new avenues of research are related to challenge 8 in Frost et al. (2015) ("How can analytical approaches keep up with advances in sequencing?"). While presenting a plethora of technical hurdles to be overcome, real-time/on-line analysis seems to be the way to deal with the ever growing mass of available data and, more importantly, provide public health authorities with useful information that can be

used in mitigation strategies. The call of Pybus et al. (2013) is now perhaps truer than ever; we must – hastily – prepare for an era of genomic plenty.

Appendix A

A pipeline for assessing convergence of MCMC for phylogenetics

BEAST_output_analysis_pipeline

Luiz Max Carvalho

25 January 2018

Preparation

First, let's specify the folder in which

```
folder <- "../examples/denv4/poor/"
ntrees <- 1000
```

Continuous parameters

The aim in this first section is to analyse the samples obtained for continuous parameters such as the evolutionary rate, (log) population sizes and the transition transversion rate parameter in the HKY model (\(\kappa\)).

Without further ado, let's load in the `.log` files and compute convergence diagnostics:

```
Logs <- getLogs(folder)
```

```
## Took 6.9  to load 3  log files
```

```
StepSize <- tail(Logs[[1]], 1)$state / ntrees
ProcessedLogs <- process_logs(Logs, burnin  = 10) ## using the "usual" 10% warm-up/burnin here
```

```
## Took 0.228  to process 3  log files
```

Here's a selection of parameters of interest:

```
ParametersOfInterest <- c("posterior", "prior",
                          "treeModel.rootHeight", "treeLength",
                          "skygrid.logPopSize1",
                          "CP1.alpha",
                          "CP1.kappa", "CP2.kappa",
                          "meanRate", "coefficientOfVariation",  "covariance")
```

and let's check whether the processed `.log` are correctly formed and ready for analysis

```
check_continuity(logs = ProcessedLogs, pars = ParametersOfInterest)
```

```
## all good, analysis can proceeed
```

```
## [1] TRUE
```

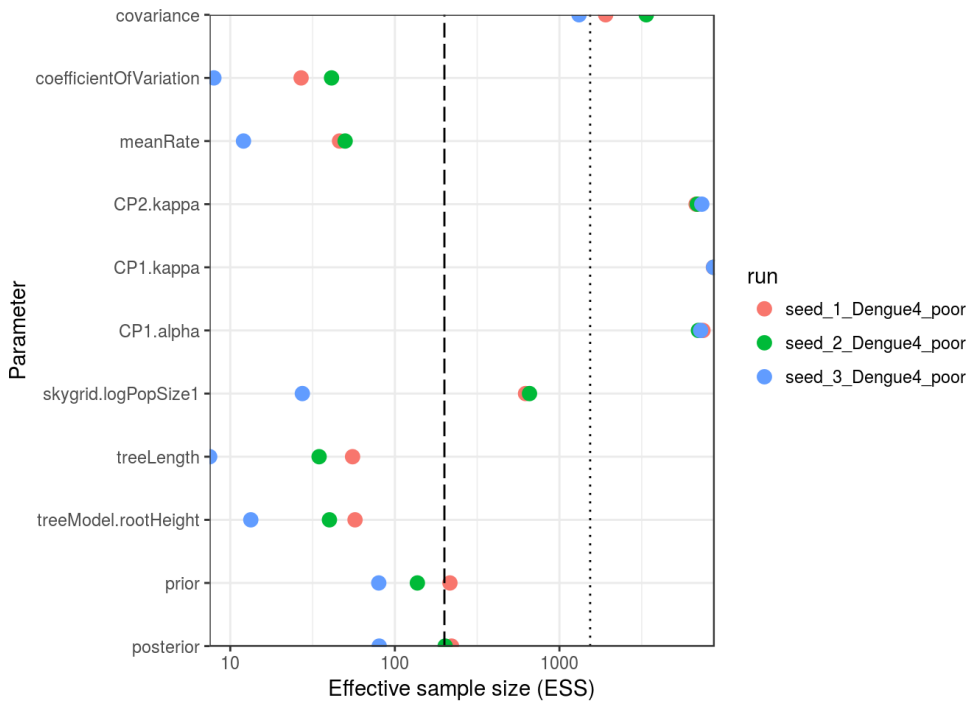
Yep. Seems so. Let's compute and look at the effective sample sizes, both univariate and multivariate:

```
( univariateESSs <- get_univariate_ESS(ProcessedLogs, pars = ParametersOfInterest) )
```

```
## Took 0.394  to compute univariate ESS for  3  log files
```

```
## $seed_1_Dengue4_poor.log
##              posterior                  prior   treeModel.rootHeight
##               220.91394              216.05858               57.27885
##             treeLength    skygrid.logPopSize1             CP1.alpha
##                55.34284              622.21883            7446.86934
##              CP1.kappa              CP2.kappa               meanRate
##             8620.01176             6763.23446               46.11481
## coefficientOfVariation             covariance
##                26.93589             1908.68995
##
## $seed_2_Dengue4_poor.log
##              posterior                  prior   treeModel.rootHeight
##               202.13019              137.00821               40.03462
##             treeLength    skygrid.logPopSize1             CP1.alpha
##                34.65403              657.72827            6992.82321
##              CP1.kappa              CP2.kappa               meanRate
##             8695.13737             6876.45783               49.86930
## coefficientOfVariation             covariance
##                41.23136             3370.53935
##
## $seed_3_Dengue4_poor.log
##              posterior                  prior   treeModel.rootHeight
##               80.332919              79.821737               13.325899
##             treeLength    skygrid.logPopSize1             CP1.alpha
##                7.496819              27.414127            7202.707626
##              CP1.kappa              CP2.kappa               meanRate
##            8640.485744            7325.688621               12.031224
## coefficientOfVariation             covariance
##                7.949814            1315.329584
```

```
ESSForPlot <- data.table::melt(
  data.table::rbindlist(
    lapply(seq_along(univariateESSs), function(i){
      x <- univariateESSs[[i]]
      res <- data.frame(
        matrix(x, nrow = 1),
        gsub(".log", "", names(univariateESSs)[i])
      )
      names(res) <- c(names(x), "run")
      return(res)
    } )
  ), id.vars = "run", variable.name = "parameter"
)
ggplot(data = ESSForPlot, aes(y = parameter, colour = run, x = value)) +
  geom_point(size = 3) +
  scale_y_discrete("Parameter", expand = c(0, 0)) +
  scale_x_log10("Effective sample size (ESS)", expand = c(0, 0)) +
  geom_vline(xintercept = 200, linetype = "longdash") +
  geom_vline(xintercept = mcmcse::minESS(p = 1, alpha = .05, eps = .1), linetype = "dotted") +
  theme_bw()
```



```
# Minimum ESS in each run
sapply(univariateESSs, summaryMin)
```

```
##                     seed_1_Dengue4_poor.log
## "coefficientOfVariation:26.9359936845359"
##                     seed_2_Dengue4_poor.log
##             "treeLength:34.6540298781855"
##                     seed_3_Dengue4_poor.log
##             "treeLength:7.49681885059039"
```

```
# Multivariate ESS (Vats et al. 2015)
mESS <- lapply(ProcessedLogs, function(x) mcmcse::multiESS(x[, ParametersOfInterest]))
unlist(mESS)
```

```
## seed_1_Dengue4_poor.log seed_2_Dengue4_poor.log seed_3_Dengue4_poor.log
##               1299.6141               1212.8027                937.0279
```

```
mcmcse::minESS(p = length(ParametersOfInterest), alpha = .05, eps = .05) ## minimum ESS needed
```

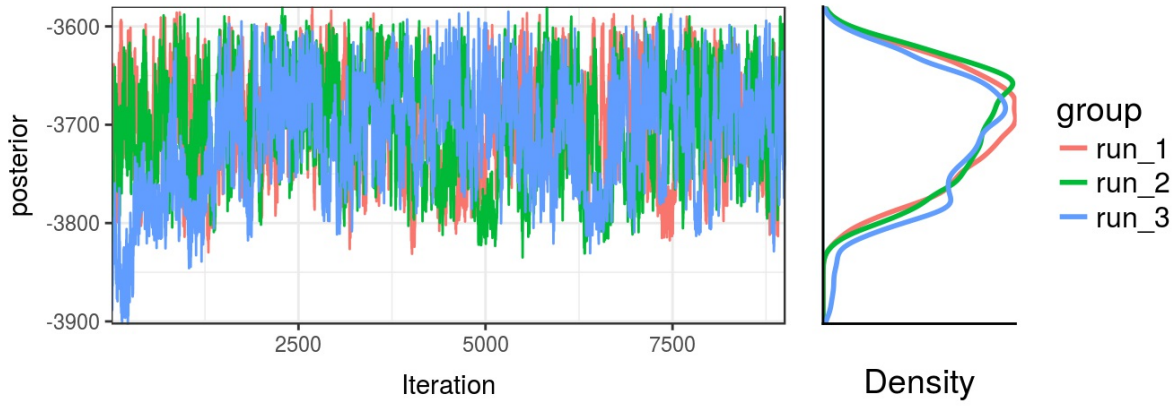
```
## minESS
##   8831
```

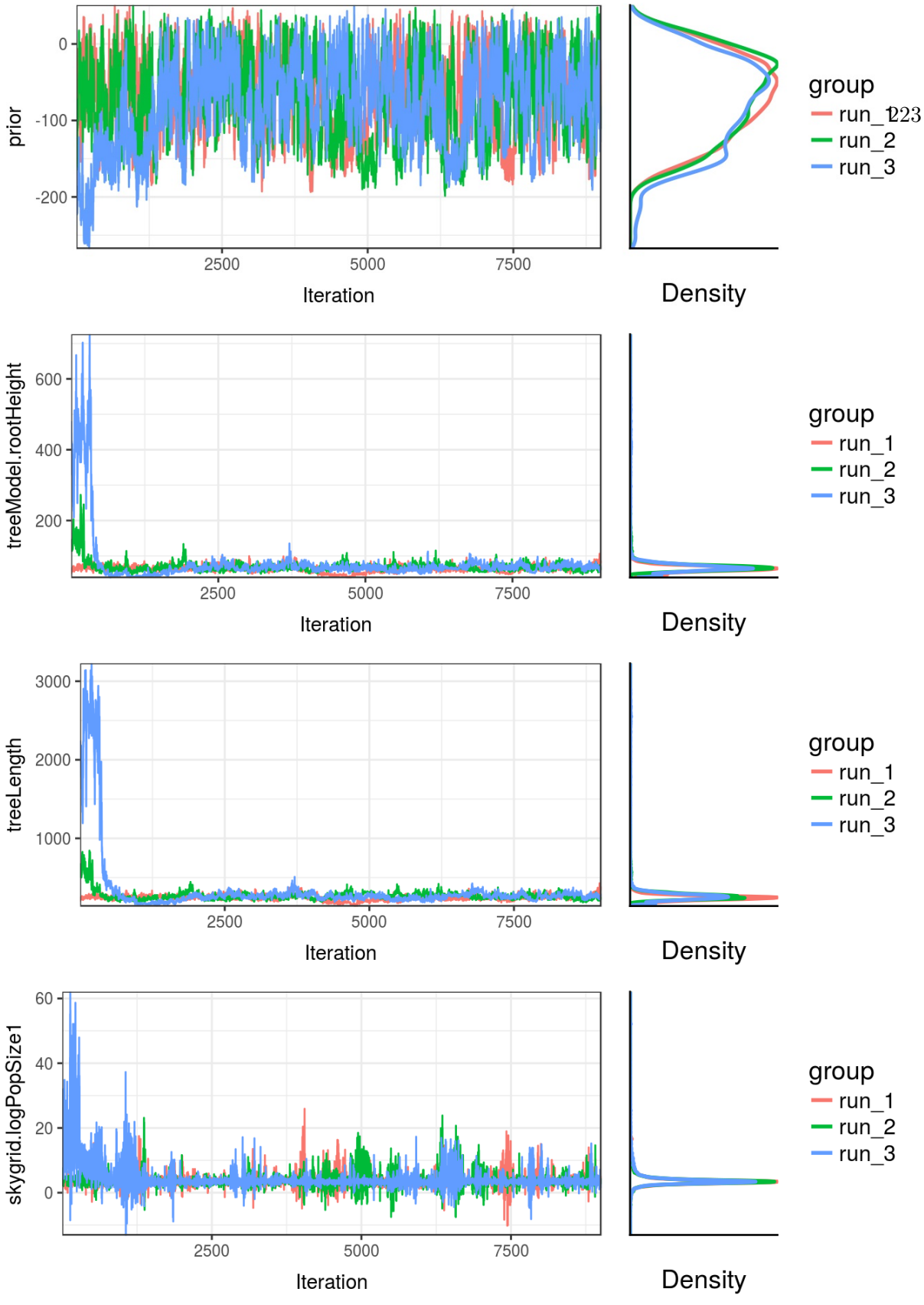
Let's look at the traces and also compute the potential scale reduction factor (PSRF) – both univariate and multivariate:

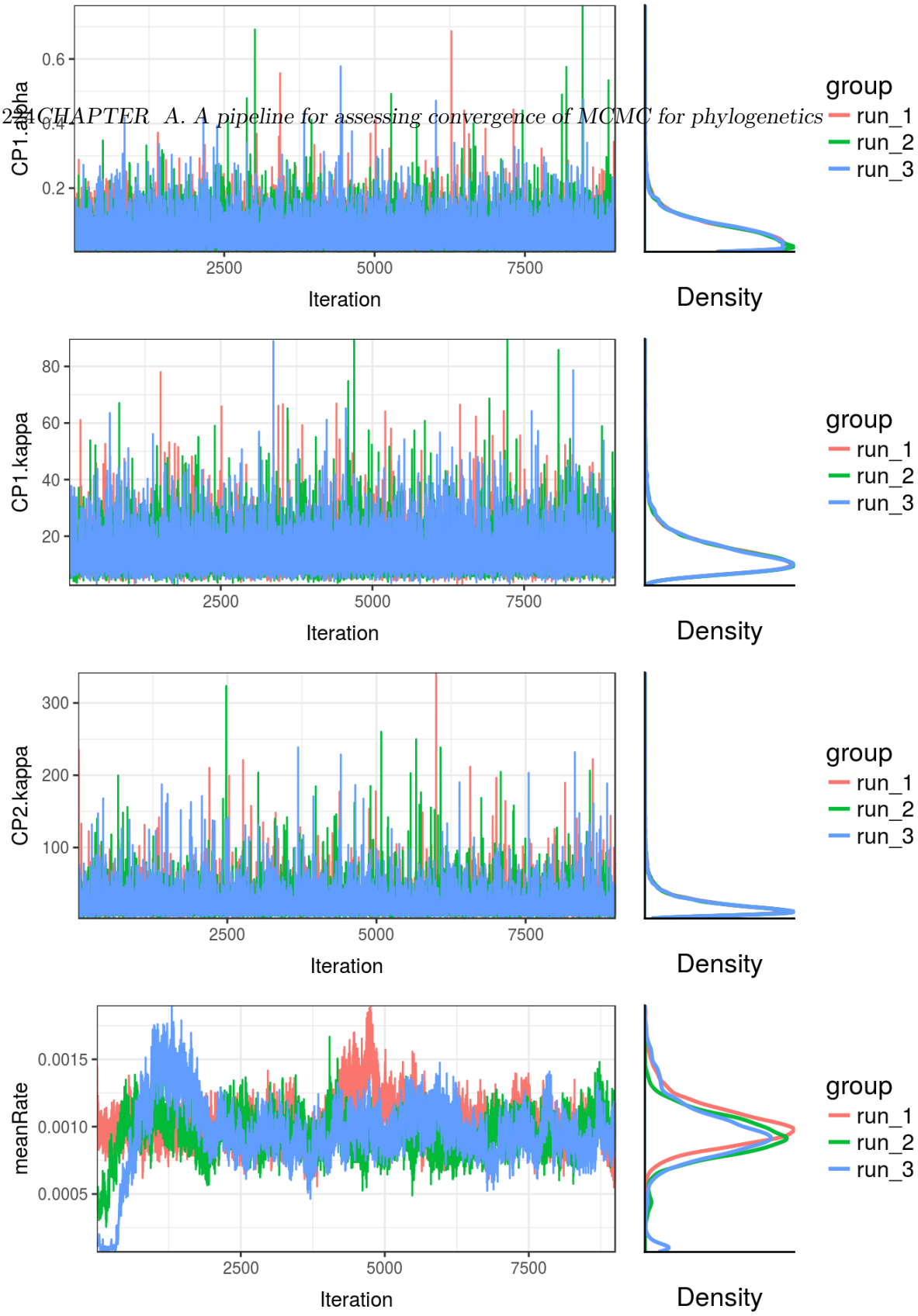
```
contPar.chain.obj <- coda::as.mcmc.list(lapply(ProcessedLogs, function(y) as.mcmc(y[, ParametersOfInterest])))
coda::gelman.diag(contPar.chain.obj)
```

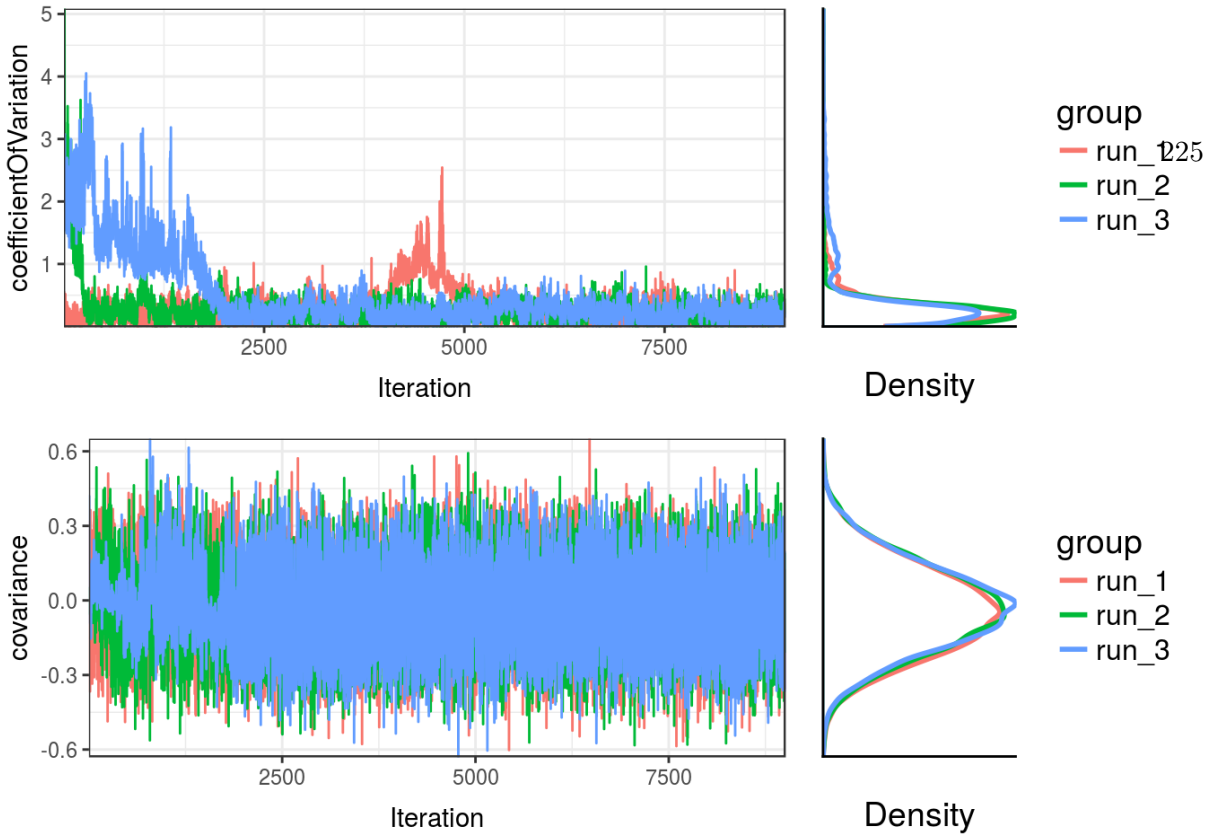
```
## Potential scale reduction factors:
##
##                         Point est. Upper C.I.
## posterior                     1.00       1.01
## prior                         1.00       1.01
## treeModel.rootHeight          1.11       1.33
## treeLength                    1.17       1.49
## skygrid.logPopSize1           1.01       1.02
## CP1.alpha                     1.00       1.00
## CP1.kappa                     1.00       1.00
## CP2.kappa                     1.00       1.00
## meanRate                      1.17       1.49
## coefficientOfVariation        1.12       1.28
## covariance                    1.00       1.00
##
## Multivariate psrf
##
## 1.13
```

```
plot_trace_new(contPar.chain.obj)
```




prior
0
-100
-200
2500
5000
7500
Iteration
Density
group
run_1
run_2
run_3
treeModel.rootHeight
600
400
200
treeLength
3000
2000
1000
skygrid.logPopSize1
60
40
20
0


CP1.alpha
0.6
0.4
0.2
2500
5000
7500
Iteration
Density
group
run_1
run_2
run_3
CP1.kappa
80
60
40
20
2500
5000
7500
Iteration
Density
group
run_1
run_2
run_3
CP2.kappa
300
200
100
2500
5000
7500
Iteration
Density
group
run_1
run_2
run_3
meanRate
0.0015
0.0010
0.0005
2500
5000
7500
Iteration
Density
group
run_1
run_2
run_3



Exploration of Phylogenetic space In this section we will explore several diagnostics measures of convergence in phylogenetic (tree) space.

Using lower-triangle matrices of distances between trees produced by `TopologyTracer`, we will compute a multidimensional scaling (MDS) representation of phylogenetic space and then use the `plotGrovesD3()` function in the **trespace** package.

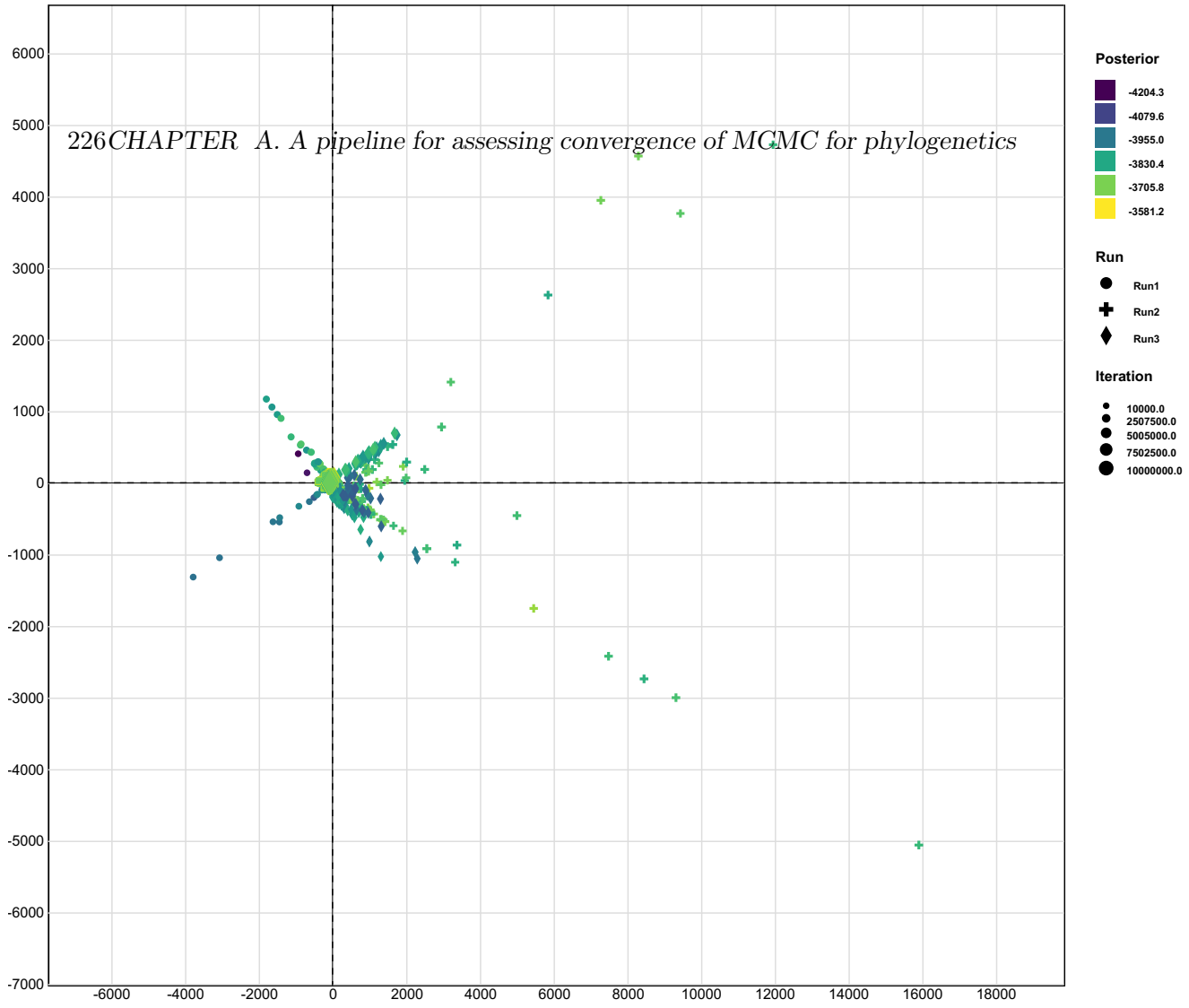
```
LT.list <- getLTs(folder) ## lower triangle matrices
```

```
## Loading required package: parallel
```

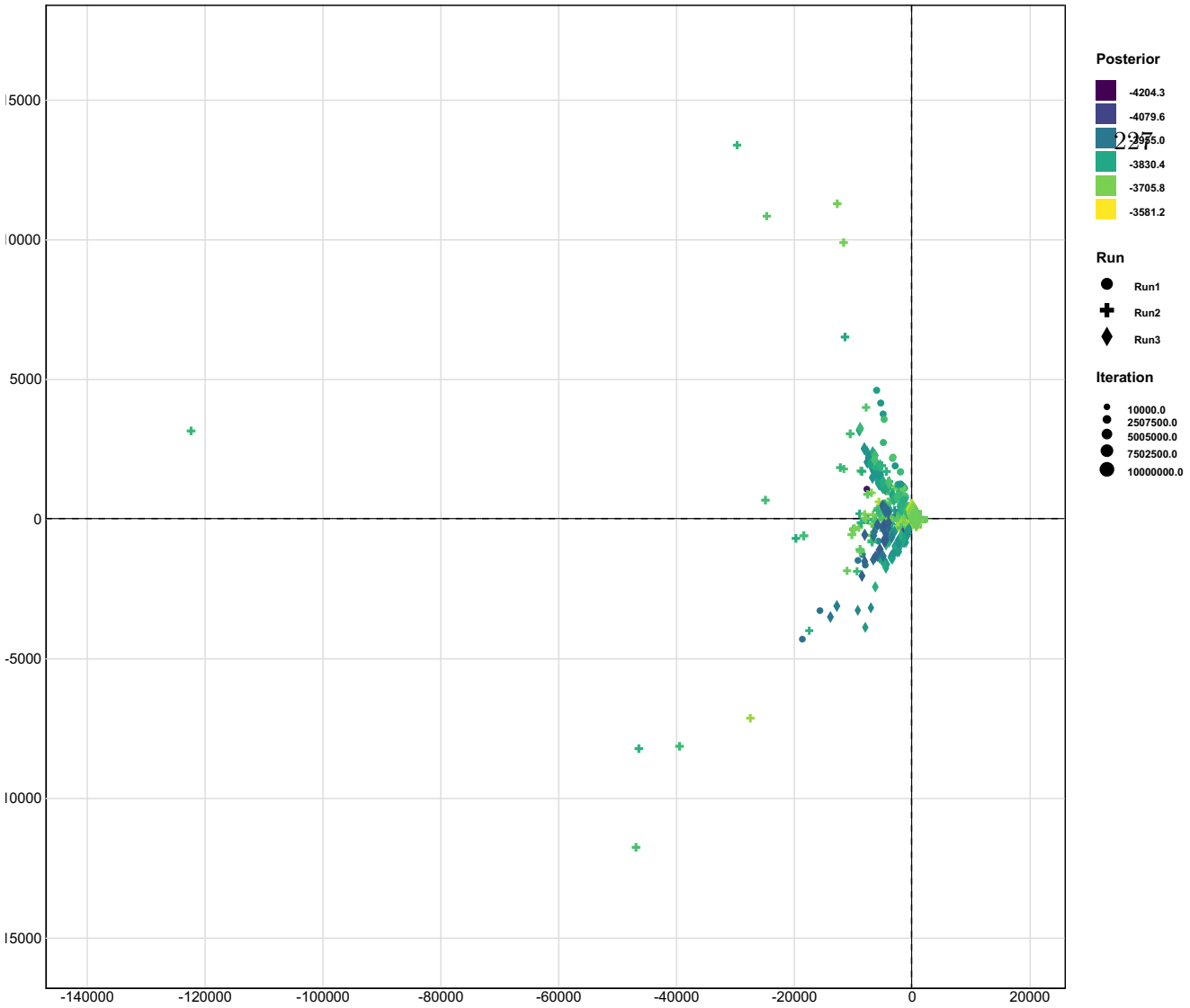
```
Full.list <- lapply(LT.list, make_full)
names(Full.list) <- gsub(".csv", "", names(Full.list))

MDS <- lapply(Full.list, getMDS, step_size = StepSize)

complete.MDS <- lapply(seq_along(MDS), function(i) {
    stem <- gsub(".csv", "", paste(strsplit(names(MDS)[i], "_")[[1]][-1], collapse = "_"))
    j <- grep(stem, names(Logs))
    TheLog <- Logs[[j]][match(seq(0, tail(Logs[[j]], 1)$state, by = tail(Logs[[j]], 1)$state / ntrees), Logs[[j]]
$state), ]
    return(
      data.frame(MDS[[i]],
                 likelihood = TheLog$likelihood,
                 posterior = TheLog$posterior)
    )
  } )
  names(complete.MDS) <- names(MDS)
## Let's grab just the distance matrices from the Kendall-Coljin metric with lambda = 1/2
MDS.Kc.half <- complete.MDS[grep("KChalf", names(complete.MDS))]
plotMDS.list(MDS.Kc.half, exclude = TRUE)
```




```
## Same now for the Steel-Penny metric
MDS.SP <- complete.MDS[grep("SP", names(complete.MDS))]
plotMDS.list(MDS.SP, exclude = TRUE)
```



Now let's look at clade frequencies

```
Clade_info <- get_clade_data(folder,  step_size = StepSize)
Clade_map_summaries <- lapply(Clade_info, function(x) summarise.clademap(x$map))
Clade.uni.ESS <- lapply(Clade_info, function(x) apply(x$map, 2, coda::effectiveSize))
lapply(Clade.uni.ESS,  function(y) mean(y <= 0)) ## proportion of clades stuck
```

```
## $seed_1_Dengue4_poor
## [1] 0.2291667
##
## $seed_2_Dengue4_poor
## [1] 0.247191
##
## $seed_3_Dengue4_poor
## [1] 0.2268041
```

```
lapply(Clade.uni.ESS,  function(y) mean(y[y > 0])) ## mean uESS (given not stuck)
```

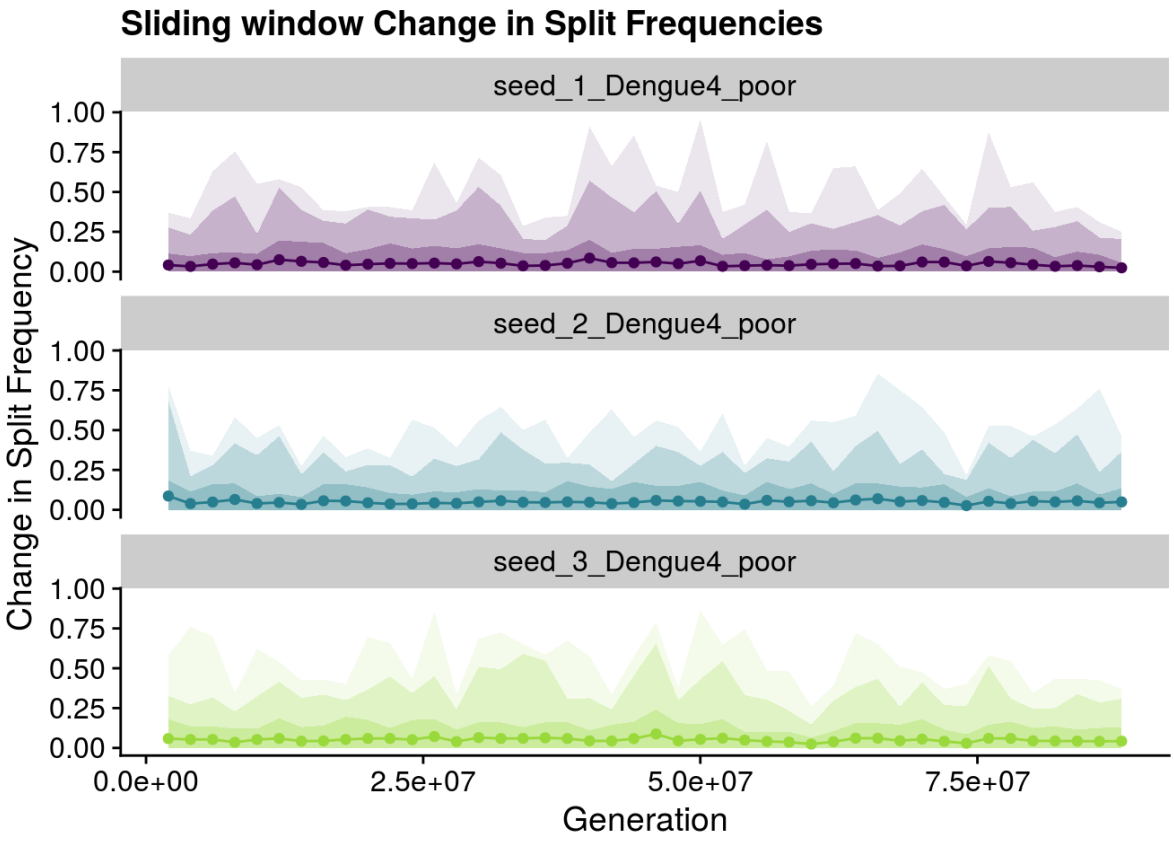
```
## $seed_1_Dengue4_poor
## [1] 874.0405
##
## $seed_2_Dengue4_poor
## [1] 917.3195
##
## $seed_3_Dengue4_poor
## [1] 837.3541
```

```
lapply(lapply(Clade_map_summaries, function(s) s$transition.nos/s$maximum), mean) ## mean clade switching score
```

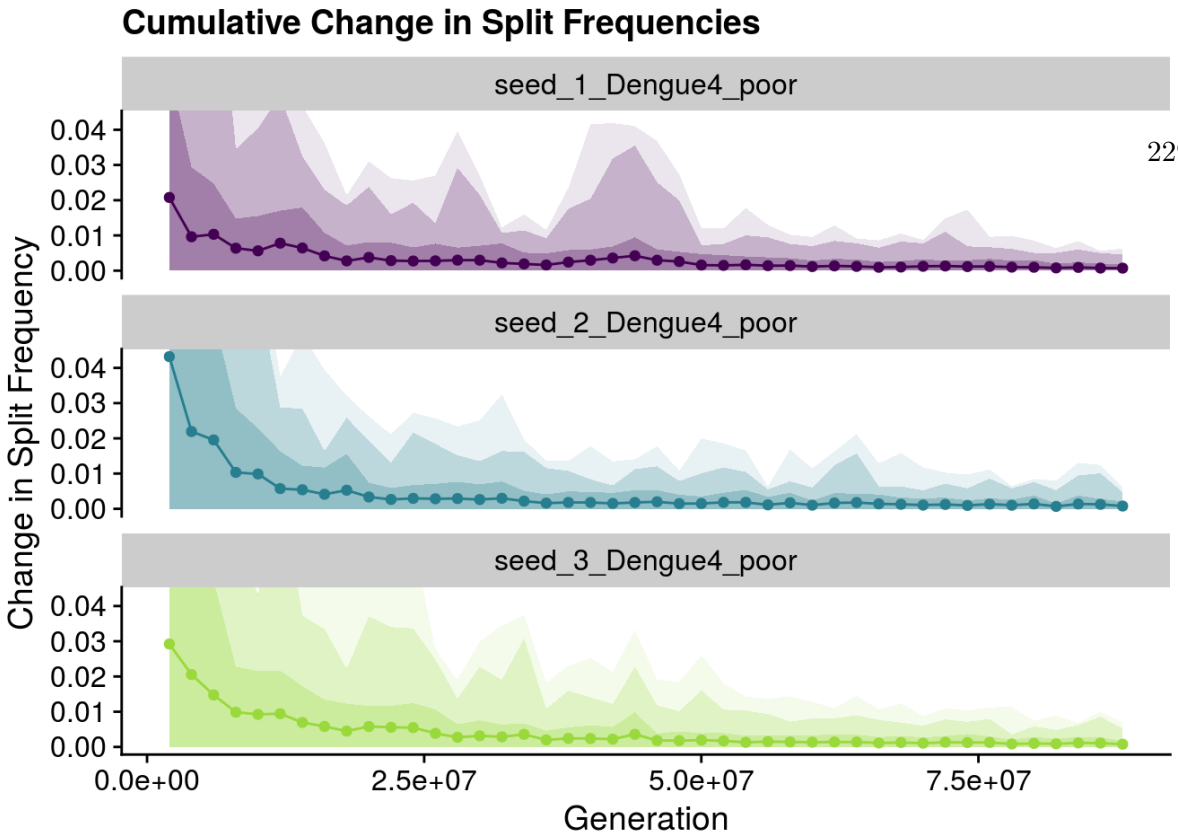
```
## $seed_1_Dengue4_poor
## [1] 0.169045
##
## $seed_2_Dengue4_poor
## [1] 0.1837223
##
## $seed_3_Dengue4_poor
## [1] 0.1606856
```



```
SlidingWCladeInfo <- lapply(Clade_info, make_fake_slidefreq, window_size = 200, step_size = StepSize)
makeplot.acsf.sliding_mod(slide.freq.list = SlidingWCladeInfo,
                          facet = TRUE)
```




```
CumulativeCladeInfo <- lapply(Clade_info, make_fake_cumfreq, window_size = 200, step_size = StepSize)
makeplot.acsf.cumulative_mod(cumulative.freq.list = CumulativeCladeInfo,
                             facet = TRUE)
```



Now let's look at clade standard deviations. Here I will be stringent and consider the maximum average standard deviation for each run. A good rule of thumb is to consider convergence is this quantity is below \(0.05\).

```
lapply(CumulativeCladeInfo, function(x) max(rwty:::get.acsf(x)$max))
```

```
## $seed_1_Dengue4_poor
## [1] 0.185
##
## $seed_2_Dengue4_poor
## [1] 0.39
##
## $seed_3_Dengue4_poor
## [1] 0.29
```

```
  topological.approx.ess_mod(mat = Full.list)
```

```
## [1] "Calculating approximate ESS with sampling intervals from 1 to 100"
```

```
##    operator approx.ess                     chain
## 1         =  141.11897     BC_seed_1_Dengue4_poor
## 2         =  158.12057     BC_seed_2_Dengue4_poor
## 3         =   21.97844     BC_seed_3_Dengue4_poor
## 4         = 1001.00000     CD_seed_1_Dengue4_poor
## 5         =  424.53399     CD_seed_2_Dengue4_poor
## 6         =   63.00512     CD_seed_3_Dengue4_poor
## 7         =   34.87053    KC0_seed_1_Dengue4_poor
## 8         =   44.95587    KC0_seed_2_Dengue4_poor
## 9         =   39.73087    KC0_seed_3_Dengue4_poor
## 10        =  171.91435    KC1_seed_1_Dengue4_poor
## 11        = 1001.00000    KC1_seed_2_Dengue4_poor
## 12        =   21.51045    KC1_seed_3_Dengue4_poor
## 13        =  185.18373 KChalf_seed_1_Dengue4_poor
## 14        = 1001.00000 KChalf_seed_2_Dengue4_poor
## 15        =   20.80479 KChalf_seed_3_Dengue4_poor
## 16        =   63.58256     RF_seed_1_Dengue4_poor
## 17        =  107.52807     RF_seed_2_Dengue4_poor
## 18        =   39.10279     RF_seed_3_Dengue4_poor
## 19        =  159.75779     SP_seed_1_Dengue4_poor
## 20        = 1001.00000     SP_seed_2_Dengue4_poor
## 21        <   13.15303     SP_seed_3_Dengue4_poor
```

Conclusion

Runs did **not** converge. From the traceplots it is clear that the mixing was poor and we obtained very low ESSs (uni and multivariate), PSRFs above 1.1. etc. MDS clearly shows runs did not explore the same space and in addition there are many low posterior trees.

Extra: bash scripts and data processing

Let's see what the contents look like – you'll need to have a similar folder structure and file naming:

```
system(paste("ls -sh", folder), intern = TRUE)
```

```
##  [1] "total 281M"
##  [2] " 9.4M BC_seed_1_Dengue4_poor.csv"
##  [3] " 9.4M BC_seed_2_Dengue4_poor.csv"
##  [4] " 9.3M BC_seed_3_Dengue4_poor.csv"
##  [5] " 9.3M CD_seed_1_Dengue4_poor.csv"
##  [6] " 9.4M CD_seed_2_Dengue4_poor.csv"
##  [7] " 9.4M CD_seed_3_Dengue4_poor.csv"
##  [8] " 1.7M clades matrix_seed_1_Dengue4_poor.cmap"
##  [9] " 1.6M clades matrix_seed_2_Dengue4_poor.cmap"
## [10] " 1.8M clades matrix_seed_3_Dengue4_poor.cmap"
## [11] "  12K cladetable_seed_1_Dengue4_poor.txt"
## [12] "  12K cladetable_seed_2_Dengue4_poor.txt"
## [13] "  12K cladetable_seed_3_Dengue4_poor.txt"
## [14] "  48K Dengue4_poor.xml"
## [15] " 8.0K full"
## [16] " 8.0K get_cmaps.sh"
## [17] " 8.0K get_distance_matrices_full.sh"
## [18] " 8.0K get_distance_matrices.sh"
## [19] " 8.0K get_equally_spaced_subsamples.sh"
## [20] " 8.9M KC0_seed_1_Dengue4_poor.csv"
## [21] " 8.9M KC0_seed_2_Dengue4_poor.csv"
## [22] " 9.0M KC0_seed_3_Dengue4_poor.csv"
## [23] " 9.3M KC1_seed_1_Dengue4_poor.csv"
## [24] " 9.3M KC1_seed_2_Dengue4_poor.csv"
## [25] " 9.3M KC1_seed_3_Dengue4_poor.csv"
## [26] " 9.3M KChalf_seed_1_Dengue4_poor.csv"
## [27] " 9.3M KChalf_seed_2_Dengue4_poor.csv"
## [28] " 9.3M KChalf_seed_3_Dengue4_poor.csv"
## [29] "1000K nohup_Dengue4_poor_1"
## [30] "1000K nohup_Dengue4_poor_2"
## [31] "1000K nohup_Dengue4_poor_3"
## [32] " 2.5M RF_seed_1_Dengue4_poor.csv"
## [33] " 2.5M RF_seed_2_Dengue4_poor.csv"
## [34] " 2.5M RF_seed_3_Dengue4_poor.csv"
## [35] " 8.0K run_beast_gnuParallel_2.0.sh"
## [36] "  16M seed_1_Dengue4_poor.log"
## [37] " 8.0K seed_1_Dengue4_poor.ops"
## [38] " 1.6M seed_1_Dengue4_poor.strees"
## [39] "  16M seed_1_Dengue4_poor.trees"
## [40] "  16M seed_2_Dengue4_poor.log"
## [41] " 8.0K seed_2_Dengue4_poor.ops"
## [42] " 1.6M seed_2_Dengue4_poor.strees"
## [43] "  16M seed_2_Dengue4_poor.trees"
## [44] "  16M seed_3_Dengue4_poor.log"
## [45] " 8.0K seed_3_Dengue4_poor.ops"
## [46] " 1.6M seed_3_Dengue4_poor.strees"
## [47] "  16M seed_3_Dengue4_poor.trees"
## [48] " 9.4M SP_seed_1_Dengue4_poor.csv"
## [49] " 9.3M SP_seed_2_Dengue4_poor.csv"
## [50] " 9.3M SP_seed_3_Dengue4_poor.csv"
```

The `.sh` files you see are used to process the `.trees` files so they can be further analysed. Most of the code featured in this section is just very simple `bash` code using the classes in BEAST to do the heavy lifting. First, let's look at the code for downsampling the `.trees` files, `get_equally_spaced_subsamples.sh`:

```
#java -Xmx4096m -cp /path/to/beast-mcmc/build/dist/beast.jar dr.app.tools.LogCombiner > /usr/bin/logcombiner
for file in *.trees
do
stem=$(basename $file .trees)
logcombiner -trees -resample 100000 -renumber $file $stem.rtrees
echo "$file is done"
done
```

Be sure to choose your `resample` argument so as to obtain ~1000 trees. For larger trees (with, say, >500 taxa) you might want to lower this to

200 trees or so, hence increase `resample`. Now we will use `TopologyTracer` to get (a) the distance of each tree to the first tree in the chain (by default, you can change the focal tree with `-tree`) and (b) a lower-triangle matrix of tree distances. We will employ the Robinson-Foulds, Steel-Penny (aka path distance) and Kendall-Coljin. The idea of computing the ESS of the distance to the focal tree was developed by Lanfear et al. (2016), who call it "pseudo-ESS". The code for `get_distance_logs.sh` is:

```
# java -Xmx4096m -cp /path/to/beast-mcmc/build/dist/beast.jar dr.app.tools.TopologyTracer > /usr/bin/treemetrics
for file in *.trees
do
stem=$(basename $file .trees)
treemetrics $file $stem.tmlog
done
```


And here are the contents of `get_distance_matrices.sh`:

```
# java -Xmx4096m -cp /home/max/beast-myfork/build/dist/beast.jar dr.app.tools.TopologyTracer > /usr/bin/treemetrics
do_distmat() {
  file=$1
  stem=$(basename $file .trees)
echo "Processing $file \n"
treemetrics -burninTrees 0 -pairwise -metric kc -lambda 0 $file KC0_$stem.csv
treemetrics -burninTrees 0 -pairwise -metric kc -lambda 0.5 $file KChalf_$stem.csv
treemetrics -burninTrees 0 -pairwise -metric kc -lambda 1 $file KC1_$stem.csv
treemetrics -burninTrees 0 -pairwise -metric sp  $file SP_$stem.csv
rm $stem.aug
}
export -f do_distmat
parallel --nice 10 --max-procs 10 do_distmat ::: *.trees
```

Appendix B

Supplementary Figures

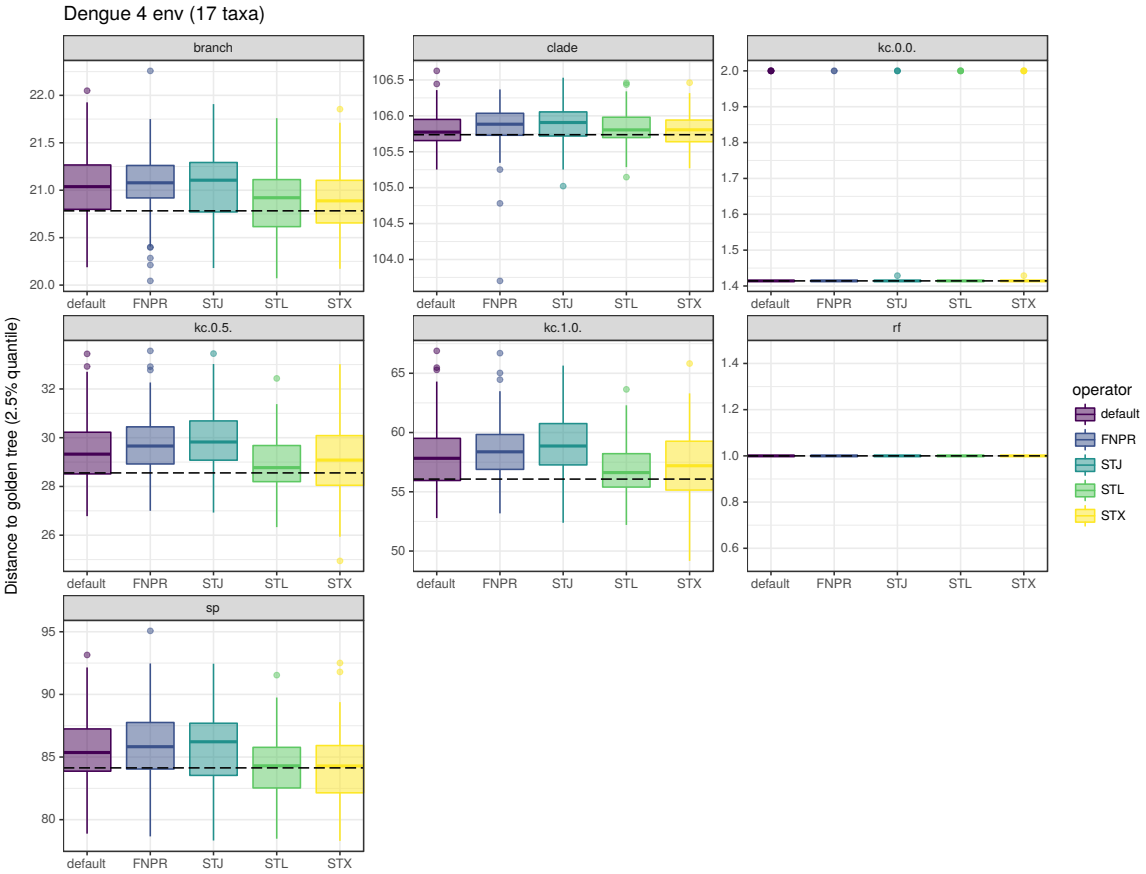


Figure B.1: Lower quantile (2.5% quantile) distance to true golden true tree for several combinations of MCMC transition kernels, Dengue 4 env data set (17 taxa).

Boxplots show the results of 100 replicates per data set. Vertical tiles show different metrics and the dashed lines show 2.5% quantiles for the target distributions shown in Figure 2.9.

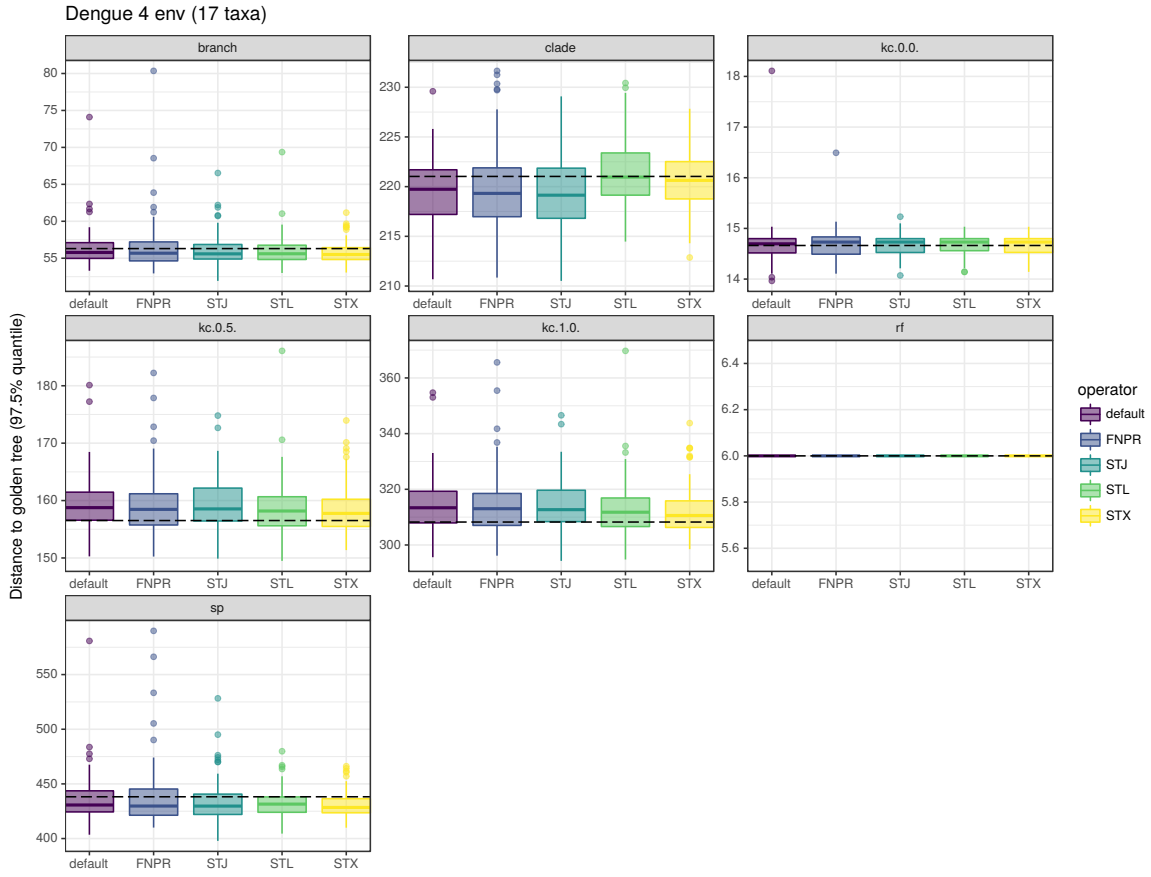


Figure B.2: Upper quantile (97.5% quantile) distance to true golden true tree for several combinations of MCMC transition kernels, Dengue 4 env data set (17 taxa).

Boxplots show the results of 100 replicates per data set. Vertical tiles show different metrics and the dashed lines show the 97.5% quantiles the target distributions shown in Figure 2.9.

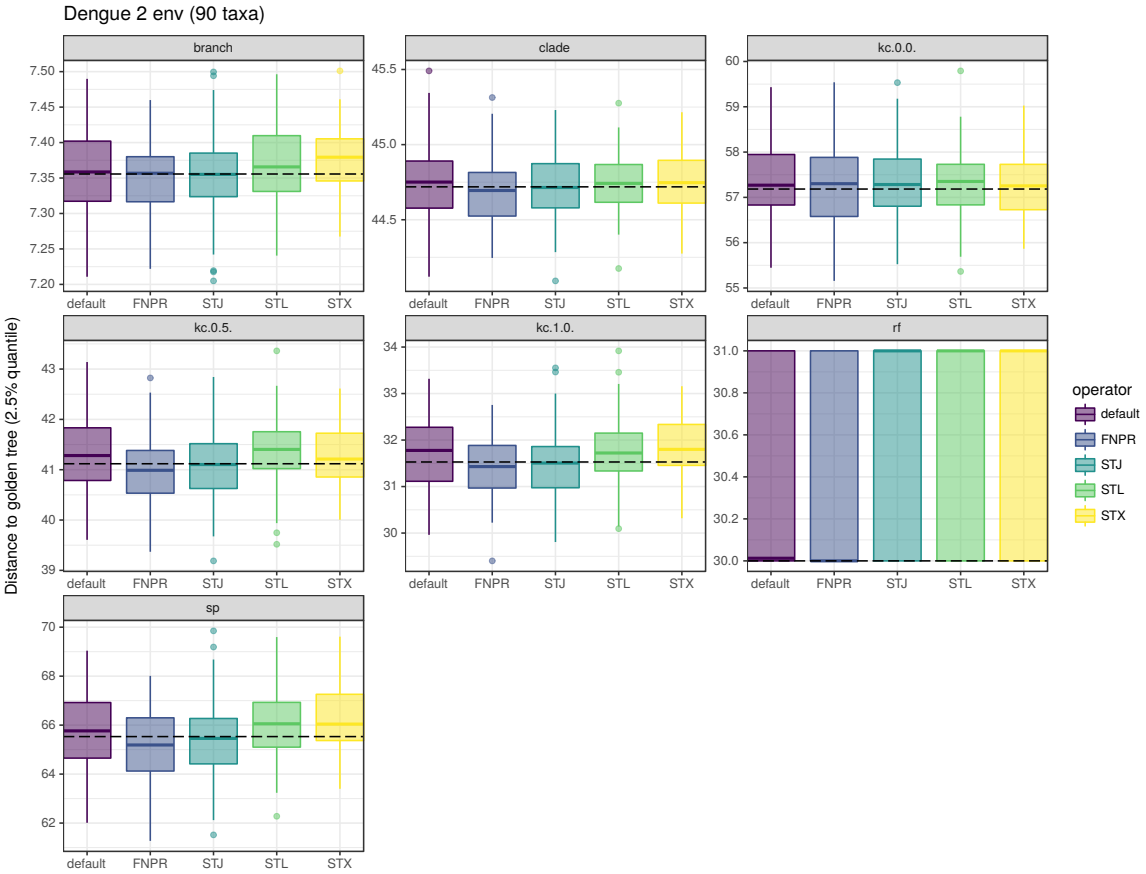


Figure B.3: Lower quantile (2.5% quantile) distance to true golden true tree for several combinations of MCMC transition kernels, Dengue 2 env data set (90 taxa).

Boxplots show the results of 100 replicates per data set. Vertical tiles show different metrics and the dashed lines show 2.5% quantiles for the target distributions shown in Figure 2.9.

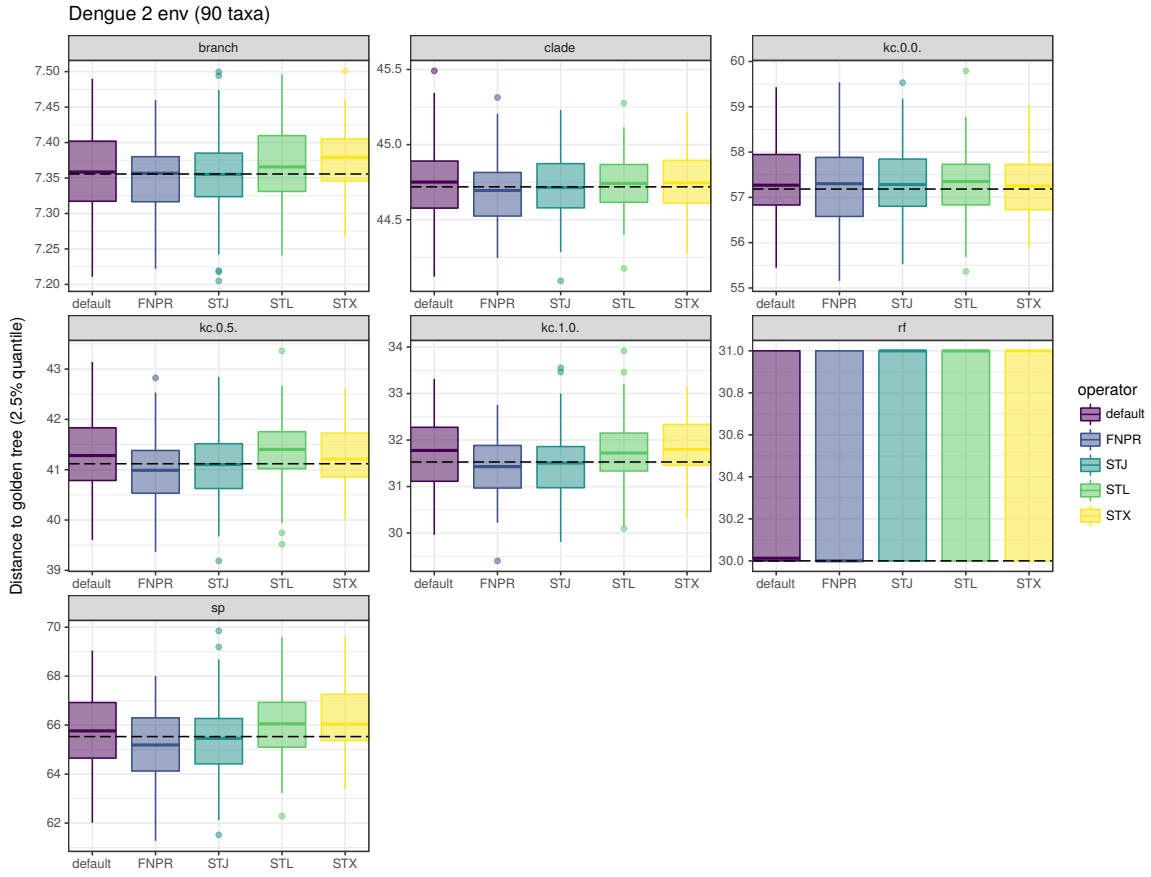


Figure B.4: Upper quantile (97.5% quantile) distance to true golden true tree for several combinations of MCMC transition kernels, Dengue 2 env data set (90 taxa).

Boxplots show the results of 100 replicates per data set. Vertical tiles show different metrics and the dashed lines show the 97.5% quantiles the target distributions shown in Figure 2.9.

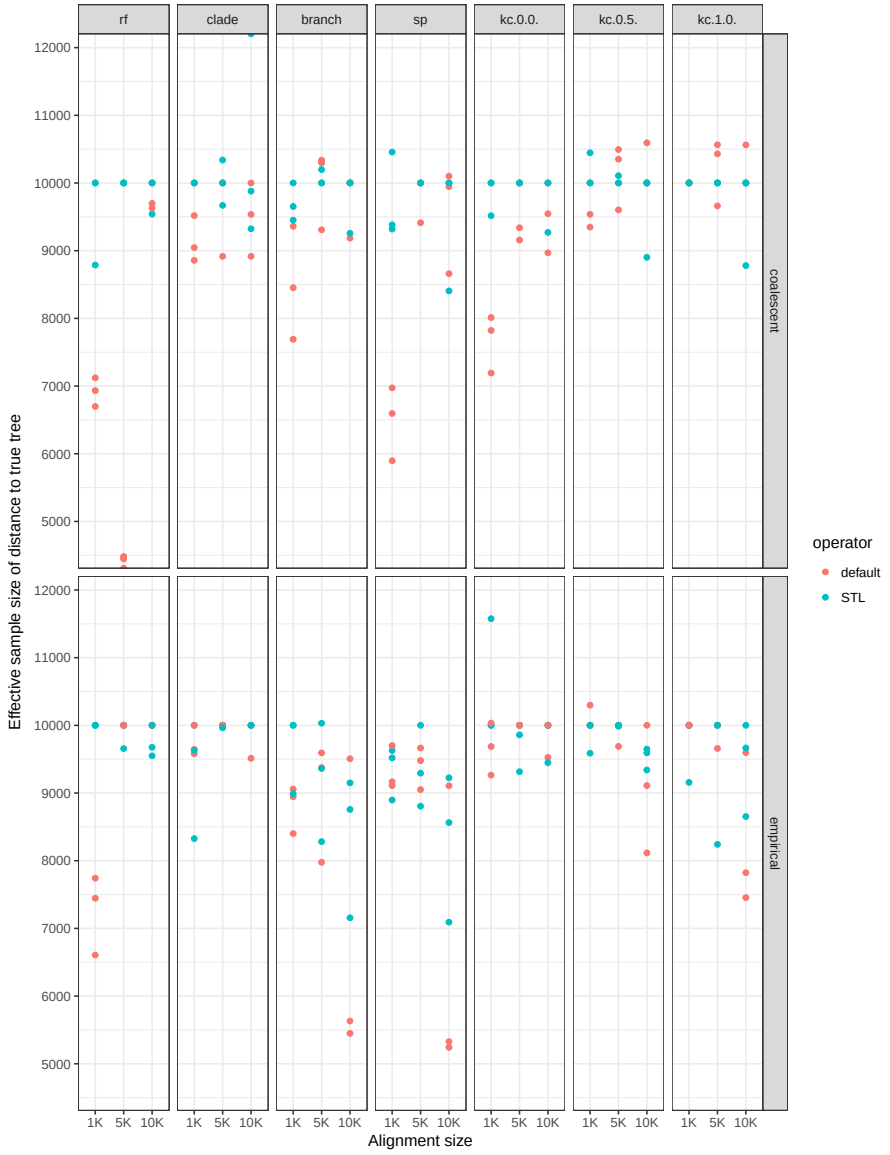


Figure B.5: Effective sample sizes of the distance to true tree for different MCMC transition kernels, simulated data sets (50 taxa).

I ran each chain for 10 million iterations, sampling trees every 1000 iterations. Each panel shows three replicates per data set. Vertical tiles show different metrics and vertical tiles show the two base trees used to simulate the data (coalescent and empirical).

References

Al-Awadhi, F., Hurn, M., and Jennison, C. (2004). Improving the acceptance rate of reversible jump mcmc proposals. *Statistics & Probability Letters*, 69(2):189–198.

Aldous, D. (1996). Probability distributions on cladograms. In *Random discrete structures*, pages 1–18. Springer.

Alexander, K. A., Sanderson, C. E., Marathe, M., Lewis, B. L., Rivers, C. M., Shaman, J., Drake, J. M., Lofgren, E., Dato, V. M., Eisenberg, M. C., et al. (2015). What factors might have led to the emergence of ebola in west africa? *PLoS neglected tropical diseases*, 9(6):e0003652.

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

Allen, B. L. and Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of combinatorics*, 5(1):1–15.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and computing*, 18(4):343–373.

Atchadé, Y. F., Rosenthal, J. S., et al. (2005). On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5):815–828.

Ayala, F. J. (1999). Molecular clock mirages. *Bioessays*, 21(1):71–75.

Baele, G., Lemey, P., Rambaut, A., and Suchard, M. A. (2017). Adaptive mcmc in Bayesian phylogenetics: an application to analyzing partitioned data in beast. *Bioinformatics (Oxford, England)*.

Baele, G., Lemey, P., and Suchard, M. A. (2015). Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. *Systematic biology*, page syv083.

Baele, G., Suchard, M. A., Rambaut, A., and Lemey, P. (2016). Emerging concepts of data integration in pathogen phylodynamics. *Systematic biology*, 66(1):e47–e65.

Baize, S., Pannetier, D., Oestereich, L., Rieger, T., Koivogui, L., Magassouba, N., Soropogui, B., Sow, M. S., Keïta, S., De Clerck, H., et al. (2014). Emergence

of zaire ebola virus disease in guinea. *New England Journal of Medicine*, 371(15):1418–1425.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Barden, D., Le, H., and Owen, M. (2018). Limiting behaviour of fréchet means in the space of phylogenetic trees. *Annals of the Institute of Statistical Mathematics*, 70(1):99–129.

Barker, D. (2015). Seeing the wood for the trees: philosophical aspects of classical, Bayesian and likelihood approaches in statistical inference and some implications for phylogenetic analysis. *Biology & Philosophy*, 30(4):505–525.

Bausch, D. G. (2017). West africa 2013 ebola: From virus outbreak to humanitarian crisis. In *Current Topics in Microbiology and Immunology*, pages 1–30. Springer.

Bayes, T. and Price, R. (1763). An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs. *Philosophical Transactions (1683-1775)*, pages 370–418.

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., et al. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217.

Bennett, S., Drummond, A., Kapan, D., Suchard, M., Munoz-Jordan, J., Pybus, O., Holmes, E., and Gubler, D. (2009). Epidemic dynamics revealed in dengue evolution. *Molecular biology and evolution*, 27(4):811–818.

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2014). GenBank. *Nucleic Acids Res.*, 42(Database issue):D32–37.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., Stuart, A., et al. (2013). Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19(5A):1501–1534.

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767.

Boehm Vock, L. F., Reich, B. J., Fuentes, M., and Dominici, F. (2015). Spatial variable selection methods for investigating acute health effects of fine particulate matter components. *Biometrics*, 71(1):167–177.

Boskova, V., Stadler, T., and Magnus, C. (2018). The influence of phylodynamic model specifications on parameter estimates of the zika virus epidemic. *Virus Evolution*, 4(1):vex044.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). Beast 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology*, 10(4):e1003537.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–39.

Brown, J. K. (1994). Probabilities of evolutionary trees. *Systematic Biology*, 43(1):78–91.

Brown, J. M. and Thomson, R. C. (2018). The behavior of metropolis-coupled Markov chains when sampling rugged phylogenetic distributions. *Systematic Biology*, page syy008.

Bryant, D. and Steel, M. (2009). Computing the distribution of a tree metric. *IEEE/ACM transactions on computational biology and bioinformatics*, 6(3):420–426.

Bryant, J. E., Holmes, E. C., and Barrett, A. D. (2007). Out of africa: a molecular perspective on the introduction of yellow fever virus into the americas. *PLoS Pathog*, 3(5):e75.

Buneman, O. P. (1971). The recovery of trees from measures of dissimilarity. *Mathematics in the archaeological and historical sciences*.

Caceres, A. J. J., Daley, S., DeJesus, J., Hintze, M., Moore, D., and John, K. S. (2011). Walks in phylogenetic treespace. *Information Processing Letters*, 111(12):600–604.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32.

Carroll, M. W., Matthews, D. A., Hiscox, J. A., Elmore, M. J., Pollakis, G., Rambaut, A., Hewson, R., García-Dorival, I., Bore, J. A., Koundouno, R., et al. (2015). Temporal and spatial analysis of the 2014-2015 ebola virus outbreak in west africa. *Nature*.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Casella, G., Robert, C. P., and Wells, M. T. (2004). Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, pages 342–347.

Centers for Disease Control (2015). Outbreaks chronology: Ebola virus disease.

Chan, K. S. and Geyer, C. J. (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1747–1758.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Cowles, M. K., Roberts, G. O., and Rosenthal, J. S. (1999). Possible biases induced by mcmc convergence diagnostics. *Journal of Statistical Computation and Simulation*, 64(1):87–104.

Cybis, G. B., Sinsheimer, J. S., Bedford, T., Mather, A. E., Lemey, P., and Suchard, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *The Annals of Applied Statistics*, 9(2):969.

Czado, C. and Raftery, A. E. (2006). Choosing the link function and accounting for link uncertainty in generalized linear models using bayes factors. *Statistical Papers*, 47(3):419–442.

Darriba, D., Flouri, T., and Stamatakis, A. (2018). The state of software for evolutionary biology. *Mol. Biol. Evol.*

Del Moral, P. (1996). Non-linear filtering: interacting particle resolution. *Markov processes and related fields*, 2(4):555–581.

Dellicour, S., Baele, G., Dudas, G., Faria, N. R., Pybus, O. G., Suchard, M. A., Rambaut, A., and Lemey, P. (2017). Phylodynamic assessment of intervention strategies for the west african ebola virus outbreak. *bioRxiv*.

Diehl, W. E., Lin, A. E., Grubaugh, N. D., Carvalho, L. M., Kim, K., Kyawe, P. P., McCauley, S. M., Donnard, E., Kucukural, A., McDonel, P., et al. (2016). Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. *Cell*, 167(4):1088–1098.

Dinh, V., Bilge, A., Zhang, C., Matsen, I., and Frederick, A. (2017). Probabilistic path hamiltonian monte carlo. *arXiv preprint arXiv:1702.07814*.

Dinh, V., Darling, A. E., Matsen, I., and Frederick, A. (2016). Online bayesian phylogenetic inference: theoretical foundations via sequential monte carlo. *Systematic biology*.

Dowle, M. and Srinivasan, A. (2017). *data.table: Extension of 'data.frame'*. R package version 1.10.4-3.

Drummond, A. (2002). *Computational Statistical Inference for Molecular Evolution and Population Genetics*. PhD thesis, The University of Auckland.

Drummond, A. J. and Bouckaert, R. R. (2015). *Bayesian evolutionary analysis with BEAST*. Cambridge University Press.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, 4(5):e88.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. (2003). Measurably evolving populations. *Trends in Ecology & Evolution*, 18(9):481–488.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7:214.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192.

Drummond, A. J. and Suchard, M. A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biol.*, 8:114.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*, 29(8):1969–1973.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, 195(2):216–222.

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D., et al. (2017). Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature*, 544(7650):309–315.

Dudas, G., Carvalho, L. M., Rambaut, A., and Bedford, T. (2018). Mers-cov spillover at the camel-human interface. *eLife*, 7.

Dudas, G. and Rambaut, A. (2014). Phylogenetic analysis of guinea 2014 ebov ebolavirus outbreak. *PLoS currents*, 6.

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*, 9(4):267–276.

Everitt, R. G., Culliford, R., Medina-Aguayo, F., and Wilson, D. J. (2018). Sequential monte carlo with transformations. *arXiv preprint arXiv:1612.06468*.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Flegal, J. M., Hughes, J., Vats, D., and Dai, N. (2017). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN. R package version 1.3-2.

Fourment, M., Claywell, B. C., Dinh, V., McCoy, C., Matsen, I., Frederick, A., and Darling, A. E. (2017). Effective online bayesian phylogenetics via sequential monte carlo with guided proposals. *Systematic biology*.

Frost, S. D., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S., and Bedford, T. (2015). Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92.

Fu, Y.-X. (2006). Exact coalescent for the wright–fisher model. *Theoretical population biology*, 69(4):385–394.

Garnier, S. (2018). *viridis: Default Color Maps from 'matplotlib'*. R package version 0.4.1.

Garthwaite, P. H., Fan, Y., and Sisson, S. A. (2016). Adaptive optimal scaling of metropolis–hastings algorithms using the robbins–monro process. *Communications in Statistics-Theory and Methods*, 45(17):5098–5111.

Gavryushkin, A. and Drummond, A. J. (2016). The space of ultrametric phylogenetic trees. *Journal of theoretical biology*, 403:197–208.

Gavryushkin, A., Whidden, C., and Matsen, F. A. (2018). The combinatorics of discrete time-trees: theory and open problems. *Journal of mathematical biology*, 76(5):1101–1121.

Gavryushkina, A., Welch, D., and Drummond, A. J. (2013). Recursive algorithms for phylogenetic tree counting. *Algorithms for Molecular Biology*, 8(1):26.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal.*, 1(3):515–534.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15):2865–2873.

Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevelhierarchical models*, volume 1. Cambridge University Press New York, NY, USA.

Gelman, A., Roberts, G. O., Gilks, W. R., et al. (1996). Efficient metropolis jumping rules. *Bayesian statistics*, 5(599-608):42.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

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

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Geyer, C. (2011). Introduction to Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo*, pages 3–48.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood.

Gill, M. S., Lemey, P., Bennett, S. N., Biek, R., and Suchard, M. A. (2016). Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. *Systematic biology*, 65(6):1041–1056.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2012). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution*, 30(3):713–724.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., Yang, X., Jiang, P. P., Nekoui, M., Colubri, A., Coomber, M. R., Fonnie, M., Moigboi, A., Gbakie, M., Kamara, F. K., Tucker, V., Konuwa, E., Saffa, S., Sellu, J., Jalloh, A. A., Kovoma, A., Koninga, J., Mustapha, I., Kargbo, K., Foday, M., Yillah, M., Kanneh, F., Robert, W., Massally, J. L., Chapman, S. B., Bochicchio, J., Murphy, C., Nusbaum, C., Young, S., Birren, B. W., Grant, D. S., Scheiffelin, J. S., Lander, E. S., Happi, C., Gevao, S. M., Gnirke, A., Rambaut, A., Garry, R. F., Khan, S. H., and Sabeti, P. C. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 179–198. Oxford University Press, Oxford.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. (2017). Nextstrain: real-time tracking of pathogen evolution. *bioRxiv*, page 224048.

Hall, M. (2015). *The phylodynamics of infectious diseases of livestock: preparing for the era of large-scale sequencing*. PhD thesis, University of Edinburgh.

Harris, T. E. (1956). The existence of stationary measures for certain markov processes. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 113–124.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2:e616.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22(2):160–174.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, 27(3):570–580.

Hillis, D. M., Heath, T. A., and St John, K. (2005). Analysis and visualization of tree space. *Syst. Biol.*, 54(3):471–482.

Hinde, J. (1982). Compound poisson regression models. In *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, pages 109–121. Springer.

Ho, L. S. T. and Ane, C. (2014). A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic Biology*, 63:397–408.

Ho, S. Y. and Duchêne, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular ecology*, 23(24):5947–5965.

Ho, S. Y., Duchêne, S., and Duchêne, D. (2015). Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Molecular ecology resources*, 15(4):688–696.

Ho, S. Y. and Larson, G. (2006). Molecular clocks: when timesare a-changin'. *TRENDS in Genetics*, 22(2):79–83.

Ho, S. Y. and Phillips, M. J. (2009). Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology*, 58(3):367–380.

Höhna, S., Defoin-Platel, M., and Drummond, A. J. (2008). Clock-constrained tree proposal operators in Bayesian phylogenetic inference. In *8th IEEE International Conference on BioInformatics and BioEngineering. BIBE 2008*, pages 1–7. IEEE.

Höhna, S. and Drummond, A. J. (2008). Evaluation of proposal distributions on clock-constrained trees in bayesian phylogenetic inference. In *Proceedings of the New Zealand Computer Science Research Student Conference 2008*.

Höhna, S. and Drummond, A. J. (2012). Guided tree topology proposals for Bayesian phylogenetic inference. *Systematic biology*, 61(1):1–11.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, 65(4):726–736.

Holder, M. and Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews genetics*, 4(4):275.

Holder, M. T., Lewis, P. O., Swofford, D. L., and Larget, B. (2005). Hastings ratio of the local proposal used in Bayesian phylogenetics. *Systematic biology*, 54(6):961–965.

Holland, B. R. (2013). The rise of statistical phylogenetics. *Australian & New Zealand Journal of Statistics*, 55(3):205–220.

Holmes, E. C., Dudas, G., Rambaut, A., and Andersen, K. G. (2016). The evolution of ebola virus: Insights from the 2013–2016 epidemic. *Nature*, 538(7624):193–200.

Huelsenbeck, J. P., Larget, B., Miller, R. E., and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic biology*, 51(5):673–688.

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

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314.

Ives, A. R. and Garland Jr, T. (2009). Phylogenetic logistic regression for binary dependent variables. *Systematic biology*, 59(1):9–26.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge university press.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). treespace: statistical exploration of landscapes of phylogenetic trees. *Molecular Ecology Resources*.

Kass, R. and Raftery, A. (1995). Bayes Factors. *J. Amer. Statist. Assoc.*, 90(430):773–795.

Kass, R. E. (2011). Statistical inference: the big picture. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 26(1):1.

Keith, J. M., Kroese, D. P., and Bryant, D. (2004). A generalized Markov sampler. *Methodology and Computing in Applied Probability*, 6(1):29–53.

Kendall, M. and Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular biology and evolution*, 33(10):2735–2743.

Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution*, 11(3):459–468.

Kuhner, M. K. and Yamato, J. (2014). Practical performance of tree comparison metrics. *Systematic biology*, 64(2):205–214.

Kuhner, M. K., Yamato, J., and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics*, 140(4):1421–1430.

Kuhner, M. K., Yamato, J., and Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149(1):429–434.

Kühnert, D., Wu, C.-H., and Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, genetics and evolution*, 11(8):1825–1841.

Kumar, S. (2005). Molecular clocks: four decades of evolution. *Nat. Rev. Genet.*, 6(8):654–662.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Lakner, C., Van Der Mark, P., Huelsenbeck, J. P., Larget, B., and Ronquist, F. (2008). Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Systematic biology*, 57(1):86–103.

Lanciotti, R. S., Gubler, D. J., and Trent, D. W. (1997). Molecular evolution and phylogeny of dengue-4 viruses. *Journal of General Virology*, 78(9):2279–2284.

Lanfear, R., Hua, X., and Warren, D. L. (2016). Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. *Genome Biology and Evolution*, 8(8):2319–2332.

Laplace, S. P. (1774). Mémoire sur la probabilité des causes par led évènemens. *Mémoires de Mathématque et Physique, Presentés à lAcadémie Royale des Sciences, par divers Savans & lûs dans ses Assemblées, Tome Sixiéme*, 66:621–56.

Larget, B. (2013). The estimation of tree posterior probabilities using conditional clade probability distributions. *Systematic biology*, 62(4):501–511.

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

Latouche, A., Guihenneuc-Jouyaux, C., Girard, C., and Hémon, D. (2007). Robustness of the bym model in absence of spatial variation in the residuals. *International journal of health geographics*, 6(1):39.

Laurent, A. M. S. (2004). *Understanding open source and free software licensing: guide to navigating licensing issues in existing & new software.* " O'Reilly Media, Inc.".

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., et al. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza h3n2. *PLoS Pathog*, 10(2):e1003932.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS computational biology*, 5(9):e1000520.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular biology and evolution*, 27(8):1877–1885.

Li, S., Pearl, D. K., and Doss, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, 95(450):493–508.

Li, X., Zai, J., Liu, H., Feng, Y., Li, F., Wei, J., Zou, S., Yuan, Z., and Shao, Y. (2016). The 2014 ebola virus outbreak in west africa highlights no evidence of rapid evolution or adaptation to humans. *Scientific reports*, 6.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):293–337.

Malsiner-Walli, G. and Wagner, H. (2016). Comparing spike and slab priors for Bayesian variable selection. *Austrian Journal of Statistics*, 40(4):241–264.

Marzi, A., Chadinah, S., Haddock, E., Feldmann, F., Arndt, N., Martellaro, C., Scott, D. P., Hanley, P. W., Nyenswah, T. G., Sow, S., et al. (2018). Recently identified mutations in the ebola virus-makona genome do not alter pathogenicity in animal models. *Cell reports*, 23(6):1806.

Matsen, F. A. (2006). A geometric approach to tree shape statistics. *Systematic biology*, 55(4):652–661.

Mau, B. and Newton, M. A. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 6(1):122–131.

Mau, B., Newton, M. A., and Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55(1):1–12.

Mayo, D. G. and Spanos, A. (2011). Error statistics. *Philosophy of statistics*, 7:152–198.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Springer US.

Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999). Mcmc convergence diagnostics: a reviewww. *Bayesian statistics*, 6:415–440.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer-Verlag.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular biology and evolution*, 25(7):1459–1471.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Möller, S., du Plessis, L., and Stadler, T. (2018). Impact of the tree prior on estimating clock rates during epidemic outbreaks. *Proceedings of the National Academy of Sciences*, page 201713314.

Morozova, O., Cohen, T., and Crawford, F. W. (2018). Risk ratios for contagious outcomes. *Journal of The Royal Society Interface*, 15(138):20170696.

Morris, M. (2017). Spatial models in stan: Intrinsic auto-regressive models for areal data.

Mossel, E. and Vigoda, E. (2005). Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–2209.

Murphy, W. J., Eizirik, E., O'brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W., et al. (2001). Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, 294(5550):2348–2351.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).

Nummelin, E. (1984). *General irreducible Markov chains and non-negative operators*. Cambridge University Press.

Nyakarahuka, L., Kankya, C., Krontveit, R., Mayer, B., Mwiine, F. N., Lutwama, J., and Skjerve, E. (2016). How severe and prevalent are ebola and marburg viruses? a systematic review and meta-analysis of the case fatality rates and seroprevalence. *BMC infectious diseases*, 16(1):708.

Nye, T. M. (2015). Convergence of random walks to brownian motion in phylogenetic tree-space. *arXiv preprint arXiv:1508.02906*.

Nylander, J. A., Wilgenbusch, J. C., Warren, D. L., and Swofford, D. L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, 24(4):581–583.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2018). *vegan: Community Ecology Package*. R package version 2.4-6.

Owen, M. and Provan, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(1):2–13.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Park, D. J., Dudas, G., Wohl, S., Goba, A., Whitmer, S. L., Andersen, K. G., Sealfon, R. S., Ladner, J. T., Kugelman, J. R., Matranga, C. B., Winnicki, S. M., Qu, J., Gire, S. K., Gladden-Young, A., Jalloh, S., Nosamiefan, D., Yozwiak, N. L., Moses, L. M., Jiang, P. P., Lin, A. E., Schaffner, S. F., Bird, B., Towner, J., Mamoh, M., Gbakie, M., Kanneh, L., Kargbo, D., Massally, J. L., Kamara, F. K., Konuwa, E., Sellu, J., Jalloh, A. A., Mustapha, I., Foday, M., Yillah, M., Erickson, B. R., Sealy, T., Blau, D., Paddock, C., Brault, A., Amman, B., Basile, J., Bearden, S., Belser, J., Bergeron, E., Campbell, S., Chakrabarti, A., Dodd, K., Flint, M., Gibbons, A., Goodman, C., Klena, J., McMullan, L., Morgan, L., Russell, B., Salzer, J., Sanchez, A., Wang, D., Jungreis, I., Tomkins-Tinch, C., Kislyuk, A., Lin, M. F., Chapman, S., MacInnis, B., Matthews, A., Bochicchio, J., Hensley, L. E., Kuhn, J. H., Nusbaum, C., Schieffelin, J. S., Birren, B. W., Forget, M., Nichol, S. T., Palacios, G. F., Ndiaye, D., Happi, C., Gevao, S. M., Vandi, M. A., Kargbo, B., Holmes, E. C., Bedford, T., Gnirke, A., Stroher, U., Rambaut, A., Garry, R. F., and Sabeti, P. C. (2015). Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*, 161(7):1516–1526.

Pattengale, N. D., Gottlieb, E. J., and Moret, B. M. (2007). Efficiently computing the robinson-foulds metric. *Journal of Computational Biology*, 14(6):724–735.

Pickett, K. M. and Randle, C. P. (2005). Strange bayes indeed: uniform topological priors imply non-uniform clade priors. *Molecular phylogenetics and evolution*, 34(1):203–211.

Piironen, J. and Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735.

Pillay, D., Herbeck, J., Cohen, M. S., de Oliveira, T., Fraser, C., Ratmann, O., Brown, A. L., Kellam, P., Pillay, D., Leigh-Brown, A., Fraser, C., Kellam, P., de Oliveira, T., Goosby, E., Hay, S., Johnson, D. A., De Cock, K. M., Dieffenbach, C., Ray, S., and Wasunna, C. (2015). PANGEA-HIV: phylogenetics for generalised epidemics in Africa. *Lancet Infect Dis*, 15(3):259–261.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125. Vienna, Austria.

Potter, C. C. and Swendsen, R. H. (2015). 0.234: The myth of a universal acceptance ratio for Monte Carlo simulations. *Physics Procedia*, 68:120–124.

Pybus, O., Drummond, A., Nakano, T., Robertson, B., and Rambaut, A. (2003). The epidemiology and iatrogenic transmission of hepatitis c virus in egypt: a bayesian coalescent approach. *Molecular biology and evolution*, 20(3):381–387.

Pybus, O. G., Fraser, C., and Rambaut, A. (2013). Evolutionary epidemiology: preparing for an age of genomic plenty. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 368(1614):20120193.

Pybus, O. G. and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.*, 10(8):540–550.

Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155(3):1429–1437.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071.

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., et al. (2016). Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4):395–399.

Rambaut, A., Drummond, A. J., Dong, X., Baele, G., and Suchard, M. A. (2018). Posterior summarisation in bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, page In press.

Rambaut, A., Lam, T. T., Carvalho, L. M., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen). *Virus Evolution*, 2(1):vew007.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615.

Rannala, B. (2002). Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.*, 51(5):754–760.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, 43(3):304–311.

Rannala, B., Zhu, T., and Yang, Z. (2012). Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Mol. Biol. Evol.*, 29(1):325–335.

Rasmussen, D. A., Ratmann, O., and Koelle, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS computational biology*, 7(8):e1002136.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: subjective recollections from incomplete data. *Statistical Science*, pages 102–115.

Robert, C. P. (2015). The metropolis-hastings algorithm. *arXiv preprint arXiv:1504.01896*.

Robert, C. P. and Casella, G. (2004). *Monte Carlo methods*. Wiley Online Library.

Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.

Roberts, G. O. and Rosenthal, J. S. (2006). Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains. *The Annals of Applied Probability*, pages 2123–2139.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Roberts, G. O., Rosenthal, J. S., et al. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367.

Roberts, G. O., Rosenthal, J. S., et al. (2004). General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71.

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

Rodrigo, A. G. and Felsenstein, J. (1999). Coalescent approaches to HIV population genetics. In Crandall, K., editor, *The Evolution of HIV*, pages 233–274. Johns Hopkins University Press.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, 61(3):539–542.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Rubinstein, R. Y. and Kroese, D. P. (2016). *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons.

Russell, C. A. and de Jong, M. D. (2017). Infectious disease management must be evolutionary. *Nature Ecology & Evolution*, 1(8):1053.

Sagulenko, P., Puller, V., and Neher, R. A. (2018). Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., and Meze-Hausken, E. (2013). A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *J R Stat Soc Ser C Appl Stat*, 62(1):85–100.

Schmidt, C. O. and Kohlmann, T. (2008). When to use the odds ratio or the relative risk? *International journal of public health*, 53(3):165–167.

Seaman III, J. W., Seaman Jr, J. W., and Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, 66(2):77–84.

Semple, C. and Steel, M. A. (2003). *Phylogenetics*, volume 24. Oxford University Press on Demand.

Sinsheimer, J. S., Lake, J. A., and Little, R. J. (1996). Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics*, pages 193–210.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

St. John, K. (2017). Review paper: The shape of phylogenetic treespace. *Systematic Biology*, 66(1):e83.

Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., et al. (2011). Estimating the basic reproductive number from viral sequence data. *Molecular biology and evolution*, 29(1):347–357.

Stadler, T. and Yang, Z. (2013). Dating phylogenies with sequentially sampled tips. *Systematic biology*, 62(5):674–688.

StanTeam (2017). Stan modeling language users guide and reference manual, version 2.17.0. *URL http://mc-stan. org*, 4.

Steel, M. (2014). Tracing evolutionary links between species. *American Mathematical Monthly*, 121(9):771–792.

Steel, M. and Pickett, K. M. (2006). On the impossibility of uniform priors on clades. *Molecular phylogenetics and evolution*, 39(2):585–586.

Steel, M. A. and Penny, D. (1993). Distributions of tree comparison metricssome new results. *Systematic biology*, 42(2):126–141.

Stigler, S. M. (1986). Memoir on the probability of the causes of events by Pierre Simon Laplace. *Statistical Science*, 1(3):364–378.

Suchar, V. A., Aziz, N., Bowe, A., Burke, A., and Wiest, M. M. (2018). An exploration of the spatiotemporal and demographic patterns of ebola virus disease epidemic in west africa using open access data sources. *Applied Geography*, 90:272–281.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evolution*, 4(1):vey016.

Suchard, M. A., Weiss, R. E., and Sinsheimer, J. S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular biology and evolution*, 18(6):1001–1013.

Takada, A., Robison, C., Goto, H., Sanchez, A., Murti, K. G., Whitt, M. A., and Kawaoka, Y. (1997). A system for functional analysis of ebola virus glycoprotein. *Proceedings of the National Academy of Sciences*, 94(26):14764–14769.

Tange, O. (2011). Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47.

Tavaré, S. (1986). *Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences*, volume 17, pages 57–86. Amer Mathematical Society.

Thawornwattana, Y., Dalquen, D., Yang, Z., et al. (2017). Designing simple and efficient Markov chain Monte Carlo proposal kernels. *Bayesian Analysis*.

Thorne, J. L., Kishino, H., and Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, 15(12):1647–1657.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728.

Tjelmeland, H. and Hegstad, B. K. (2001). Mode jumping proposals in mcmc. *Scandinavian Journal of Statistics*, 28(1):205–223.

To, T.-H., Jung, M., Lycett, S., and Gascuel, O. (2015). Fast dating using least-squares criteria and algorithms. *Systematic biology*, 65(1):82–97.

Urbanowicz, R. A., McClure, C. P., Sakuntabhai, A., Sall, A. A., Kobinger, G., Muller, M. A., Holmes, E. C., Rey, F. A., Simon-Loriere, E., and Ball, J. K. (2016). Human Adaptation of Ebola Virus during the West African Outbreak. *Cell*, 167(4):1079–1087.

Vats, D., Flegal, J. M., and Jones, G. L. (2015). Multivariate output analysis for Markov chain Monte Carlo. *arXiv preprint arXiv:1512.07713*.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Volz, E. and Frost, S. (2017). Scalable relaxed clock phylogenetic dating. *Virus Evolution*, 3(2).

Volz, E. M., Koelle, K., and Bedford, T. (2013). Viral phylodynamics. *PLoS Comput. Biol.*, 9(3):e1002947.

Volz, E. M., Pond, S. L. K., Ward, M. J., Brown, A. J. L., and Frost, S. D. (2009). Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430.

Vrancken, B., Lemey, P., Rambaut, A., Bedford, T., Longdon, B., Günthard, H. F., and Suchard, M. A. (2015). Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: applications to viral and host phenotypic evolution. *Methods in ecology and evolution*, 6(1):67–82.

Wang, Y. and Yang, Z. (2014). Priors in Bayesian phylogenetics. In M.-H. Chen, L. K. and Lewis, P., editors, *Bayesian Phylogenetics: Methods, Algorithms, and Applications*. Chapman & Hall/CRC, London.

Warren, D. L., Geneva, A. J., and Lanfear, R. (2017). Rwty (r we there yet): An r package for examining convergence of Bayesian phylogenetic analyses. *Molecular biology and evolution*, 34(4):1016–1020.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Welch, J. J. and Bromham, L. (2005). Molecular dating when rates vary. *Trends in Ecology & Evolution*, 20(6):320–327.

Whidden, C. and Matsen, F. A. (2015). Quantifying mcmc exploration of phylogenetic tree space. *Systematic biology*, page syv006.

Whidden, C. and Matsen, F. A. (2017). Ricci–ollivier curvature of the rooted phylogenetic subtree–prune–regraft graph. *Theoretical Computer Science*.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Willis, A. and Bell, R. (2017). Uncertainty in phylogenetic tree estimates. *Journal of Computational and Graphical Statistics*, (just-accepted).

Woolhouse, M. E. (2002). Population biology of emerging and re-emerging pathogens. *Trends in microbiology*, 10(10):s3–s7.

World Health Organization (2016a). 2014 ebola outbreak in west africa - case counts.

World Health Organization (2016b). Ebola situation report - 10 june 2016.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol. (Amst.)*, 11(9):367–372.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using dna sequences: a Markov chain Monte Carlo method. *Molecular biology and evolution*, 14(7):717–724.

Yang, Z. and Rannala, B. (2005). Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology*, 54(3):455–470.

Yang, Z. and Rodríguez, C. E. (2013). Searching for efficient Markov chain Monte Carlo proposal kernels. *Proceedings of the National Academy of Sciences*, 110(48):19307–19312.

Yoder, A. D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.*, 17(7):1081–1090.

Zanella, G. (2017). Informed proposals for local mcmc in discrete spaces. *arXiv preprint arXiv:1711.07424*.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

Zhu, S., Degnan, J. H., and Steel, M. (2011). Clades, clans, and reciprocal monophyly under neutral evolutionary models. *Theoretical Population Biology*, 79(4):220–227.

Zlateva, K. T., Lemey, P., Vandamme, A.-M., and Van Ranst, M. (2004). Molecular evolution and circulation patterns of human respiratory syncytial virus subgroup a: positively selected sites in the attachment g glycoprotein. *Journal of virology*, 78(9):4675–4683.

Zuckerkandl, E. and Pauling, L. (1962). Molecular disease, evolution and genetic heterogeneity.