Bayesian statistical modelling of genetic sequence evolution

Konstantinos Angelis

University College London

Department of Genetics, Evolution and Environment

2015

Thesis submitted to the University College London
for the degree of Doctor of Philosophy

I, Konstantinos Angelis, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. Specifically, four chapters of this thesis involve joint work which is specified below:

Chapter 3. This chapter is a modification of the work of Angelis et al. (2014) and includes contributions by Dr. Mario dos Reis (MdR) and Prof. Ziheng Yang (ZY). Konstantinos Angelis (KA), MdR and ZY designed the study, build the model and interpreted the results. KA wrote the software and performed the simulation and real data analysis.

Chapter 4. This chapter is a modification of the work of Angelis and dos Reis (2015) and includes contribution of MdR. KA and MdR designed the study and interpreted the results. KA performed the simulation and real data analyses.

Chapter 5. This chapter includes contribution of MdR and ZY. KA, MdR and ZY designed the study. KA performed the analysis and interpreted the results.

Chapter 6. This chapter is a modification of the work of dos Reis et al. (2015) and includes contribution of MdR, Yuttapong Thawornwattana (YT), Prof. Maximilian Telford, Prof. Philip Donoghue and ZY. YT prepared the data and set up the analysis pipeline. KA, YT and MdR performed the analysis. All researchers contributed to the interpretation of the results. This work is a highly collaborative effort of many people and integrates expertise from different fields such as zoology, palaeontology, molecular evolution and statistics.

In addition to the above, during my PhD studies I published the following papers, which do not form part of the thesis:

Paraskevis D, Angelis K, Magiorkinis G, Kostaki E, Ho S, Hatzakis A. 2015. Dating the origin of hepatitis B virus reveals higher substitution rate and viral adaptation to the branch leading to F/H genotypes. Mol Phylogenet Evol 93: 44-54.

Angelis K, Albert J, Mamais I (>10 authors), et al. 2014. Global dispersal pattern of HIV-1 CRF01_AE: A genetic trace of human mobility related to heterosexual activities centralized in South-East Asia. J Infect Dis 211:1735–1744.

Abstract

Bayesian statistics has been at the heart of phylogenetic inference over the last decade, particularly after the development of powerful programs that implement efficient Markov chain Monte Carlo algorithms, allowing inference from multi-parametric problems in realistic time frames. In this thesis we develop and test Bayesian methods to analyse molecular sequence data to address important biological questions. First, we review some fundamental aspects of Bayesian inference and highlight current Bayesian applications in molecular evolution with particular focus in studying natural selection and estimating species divergence times. Then, we develop a new Bayesian method to estimate the nonsynonymous/synonymous rate ratio and evolutionary distance for pairwise sequence comparisons. The new method addresses weaknesses of previous counting and maximum-likelihood methods. It is also computationally efficient and thus suitable for genome-scale screening. Then, we explore the performance of existing Bayesian algorithms in estimating species divergence times. In particular, we study the impact of ancestral population size and incomplete lineage sorting on Bayesian estimates of species divergence times under the molecular clock, when those factors of molecular evolution are ignored by the inference model. The estimates can be highly biased, especially in the case of shallow phylogenies with large ancestral population sizes. Then, using computer simulations and real data analyses we study the effect of five commonly used partitioning strategies for divergence times estimation and show that the choice of the partitioning scheme is important in case of serious clock-violation with incorrect prior assumptions. Finally, a Bayesian molecular clock dating study is performed to estimate the timeline of animal evolution. The results indicate that the time estimates are highly variable, precluding the inference of a precise timescale of animal evolution based on the current data and methods.

Acknowledgements

First and foremost, I would like to thank deeply my supervisor Ziheng Yang for his support, guidance, knowledge and useful advice he has given me throughout the three years of my PhD studies. He has been an inspiring supervisor and a perfect collaborator.

I would also like to express my sincere gratitude to my exceptional collaborator Mario dos Reis for all his guidance and valuable advice he has given me. I thank him for our perfect collaboration and for the experience and precious knowledge he gave me. He will always be a good friend.

Special thanks to Yuttapong Thawornwattana. He started the work of chapter 6 as part of his undergraduate project at UCL. I took on the analytical part of the project, and helped completed it in high collaboration with Maximilian Telford, Philip Donoghue, Mario dos Reis and Ziheng Yang.

Many thanks to all people I have worked with including the members of our lab Jose Antonio Barba Montoya and Daniel Dalquen for useful scientific feedback and discussions concerning my projects.

I would also like to particularly acknowledge the University College London and the Department of Genetics, Evolution and Environment for the financial support without which this work would not have been possible.

Lastly, I would like to thank my wife Foteini Kosmidi for her constant support and patience all these years.

Contents

List of Figures

List of Tables

Introduction

Bayesian statistics developed substantially during the late 20th century and is nowadays used to address important problems in various scientific fields. An advantage of the Bayesian framework over the classical statistics is that it integrates in a straightforward way prior information about model parameters with information from the data through the likelihood function. This property and the development of efficient Markov chain Monte Carlo algorithms, have allowed Bayesian inference to be applied to complex multi-parameter problems, thus finding applications in real life problems. Bayesian inference is used in molecular evolution and phylogenetics since the mid-90s with a number of applications such as estimation of phylogenetic trees, species divergence times, molecular evolutionary rates, demographic population histories and natural selection.

The work presented in this thesis concerns the development of new and use of existing Bayesian methods to study two important aspects of molecular evolution: natural selection and species divergence times.

In chapter 1 we give an overview of the Bayesian methodology with emphasis on some key points which are important for the better understanding of the applications described in the following chapters.

In chapter 2 we present some current applications of Bayesian statistics in phylogenetic inference with particular focus on studying natural selection and estimating species diversification times. We describe codon models and Bayesian techniques to identify positive selection and we introduce the molecular clock notion and a Bayesian molecular clock dating method to estimate species divergence times.

In chapter 3 we develop a new Bayesian method to measure selection for pairwise sequence comparisons. The new method addresses weaknesses of previous counting and maximum likelihood methods. It is computationally efficient, and can be used in genome-scale analysis of protein-coding gene sequences.

In chapter 4 we study the impact of ancestral population size and incomplete lineage sorting on Bayesian estimates of species divergence times and rate under the molecular clock, when the inference model ignores those aspects of molecular evolution. We use a combination of mathematical analysis, computer simulation, and real data analysis to study the problem. We show that the time and rate estimates can be strongly biased, particularly in case of shallow phylogenies with large population sizes.

In chapter 5 we evaluate the performance of five commonly used data partitioning strategies for the Bayesian estimation of species divergence times. We use computer simulation and real data analysis of a plant data set and we show that differences in time

estimates among the strategies are small when the priors are correct but can be large when model assumptions are violated. Especially when the clock is seriously violated and an improper clock model is used, the differences can be quite dramatic.

In chapter 6 we perform a molecular clock dating study to estimate the timeline of animal evolution. We analyze a large amino acid alignment of 54 metazoan species in combination with 34 fossil calibrations to obtain Bayesian estimates of metazoan divergence times. We explore several sources that cause uncertainties in the estimated times such as different interpretations of the fossil record, rate variation among lineages, limited molecular data, unresolved phylogenetic relationships and show that their cumulative impact precludes a precise estimation of the metazoan timescale with current data and methods.

Chapter 7 is a summary of the work.

1 Bayesian Theory

1.1 Bayesian inference

There are two main approaches for statistical data analysis: the *Frequentistic* or classical approach and the Bayesian. Bayesian ideas were introduced by Thomas Bayes during the 18th century (Stigler 1986). However, it was the French mathematician Pierre-Simon Laplace who developed further the ideas of Bayes to become what is known today as Bayesian statistics (Stigler 1986). In both approaches, there is a parameter θ which we want to estimate and a mechanism $f(x|\theta)$ which determines the probability to observe data x given a value of the parameter. The fundamental difference of the Bayesian framework from the classical approach is that the parameter θ is treated as a random variable and thus has a distribution, while in classical statistics it is considered to be an unknown constant. Although this difference might not seem that important, it leads to a markedly different statistical modelling and interpretation.

Inference in classical statistics is based on the likelihood, that is the probability of observing the data x given the value for parameter θ, $f(x|\theta)$. The value that maximizes the likelihood function is an estimate of θ. In contrast, Bayesian inference is based on the posterior distribution of θ, that is, the probability distribution of the parameter given the data $f(\theta \mid x)$. To estimate the posterior distribution, we need to specify a prior distribution $f(\theta)$, which expresses one's knowledge on θ before observing any data. Then the posterior distribution is given by the Bayes' theorem

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)}, \tag{1.1}$$

where $f(x)$ is the marginal likelihood, a constant which guarantees that the posterior distribution integrates to one and is given by $f(x) = \int f(x \mid \theta) f(\theta) \mathrm{d}\theta$. Here, θ is assumed to be continuous but, if θ is discrete the integral is substituted by a sum.

In classical statistics, since the unknown parameter θ is assumed constant, we always get point estimates and we use confidence intervals to express uncertainty around the estimated values. In the Bayesian framework the inference is the posterior distribution. However, in most cases we need to summarize the information included in the posterior into a single "best" estimate. Such point estimates could be the mean, mode or median of the posterior distribution. The analogue of the confidence interval in the Bayesian framework is the credibility interval, (c, d), which is defined as $\int_c^d f(\theta \mid x) d\theta = 1 - \alpha$ and means that the true

parameter θ is in the interval (c, d) with probability $1-\alpha$. For example, one can build a 95% equal-tail credibility interval (CI) using the 2.5% and 97.5% quantiles of the posterior distribution. However, when the posterior density is multimodal or skewed this interval may include less plausible values of θ than values outside the interval. So, in the above definition we impose the constraint that the width of the interval should be as small as possible, forming the highest posterior density (HPD) interval. Clearly the HPD interval offers two advantages over the equal-tail CI: (i) any point within the HPD interval has higher density than any point outside the interval and (ii) given a probability level $1-\alpha$ the HPD interval has the smallest width. When the posterior density is unimodal and symmetrical the HPD and equal-tail intervals are identical.

Most statistical problems involve models with more than one unknown parameters. Our interest might be to one of them or to a subgroup of them, but usually the values of the other parameters (called nuisance parameters) might be unknown. Dealing with nuisance parameters in classical statistics is hard but the Bayesian framework provides a straightforward approach to the problem. In case of multi-parameter problems we have a vector $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)$ of parameters which we want to make inference about. We specify a multivariate prior $f(\boldsymbol{\theta})$ and along with the likelihood function $f(x \mid \boldsymbol{\theta})$ we estimate the posterior using the Bayes' theorem

$$f(\boldsymbol{\theta} \mid x) = \frac{f(x \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{\int f(x \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}. \tag{1.2}$$

The posterior $f(\boldsymbol{\theta} \mid x)$ is a multivariate distribution and we can make inference about any subset of parameters by applying straightforward probability calculations. For example, the marginal posterior distribution for the parameter θ_1 can be calculated by integrating out all other parameters,

$$f(\theta_1 \mid x) = \int f(\boldsymbol{\theta} \mid x) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_p \tag{1.3}$$

Although there is no need for a new theory in multi-parameter problems there are two major practical problems caused by the increase in dimensionality. First, there is an increased difficulty in specifying the prior distribution. The prior is multidimensional and except for expressing one's beliefs for any parameter individually, should also contain information on the correlations among the parameters. This is substantially more complicated than specifying a univariate prior for a single parameter. The second problem is computational. In multivariate models the marginal likelihood involves the calculation of a multidimensional integral which might be impossible to calculate analytically or even with advanced numerical integration techniques. In this case, the Bayesian inference has been made possible by development of efficient Markov chain Monte Carlo algorithms which simulate from the posterior.

1.2 Some advantages of Bayesian statistics

An advantage of Bayesian statistics over the frequentistic approach is that it offers a straightforward interpretation. The cornerstone of classical statistics that the unknown parameter θ is being treated as constant leads to problems in interpretation. In classical statistics we'd like a 95% confidence interval [c, d] to mean that the true θ is between c and d with probability 95%, however, this is not the case. Since θ is assumed constant either is or not within the interval, and cannot lie within it with some probability. The random element is the data, and so the correct interpretation is that if we take many samples from the population and construct a confidence interval from each sample, then 95% of them will contain the true parameter value θ. In contrast, the Bayesian framework offers a straightforward interpretation; given the data the 95% credibility interval contains the true value of the parameter θ with probability 95%.

Another problem with the classical approach is the violation of the likelihood principle. The likelihood principle states that if two experiments have the same likelihood (up to proportionality) then the inference around the parameter θ should be the same. In other words the likelihood function contains all information in the data about the parameter θ and the inference should be the same from two experiments with the same likelihood. However, this might not always be the case (see example below). In contrast, in the Bayesian framework when the likelihoods from two different experiments are the same (up to proportionality) the posterior distributions will be the same, leading to the same inference. To clarify that we consider the following example proposed by Lindley and Phillips (1976). Let's assume an experiment in which we count the number of successes x in n independent trials. We are interested to test whether the probability of success (θ) in one trial is ½ or lower. In other words we want to test the null hypothesis H_0: $\theta = ½$ against the alternative H_1: $\theta < ½$. The number of positive outcomes x follows a Binomial distribution

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x},\ x = 0, 1, \ldots, n. \tag{1.4}$$

Let's assume that we observe $x = 3$ successes in $n = 12$ trials. Then the p-value, the probability to observe at most as many successes under the H_0 as in the observed data is $p_1 = P(x \leq 3 \mid \theta = 0.5) = 0.073$, meaning that the H_0 is not rejected at significance level $\alpha = 5\%$. Now let's consider an alternative experimental design in which we count the number of trials y until we observe m successes. We assume that we observe the same data, $m = 3$ successes in $y = 12$ trials. The number of trials y follows a Negative Binomial distribution

$$f(y \mid \theta) = \binom{y-1}{m-1} \theta^m (1-\theta)^{y-m},\ y = \mathrm{m}, \mathrm{m}+1, \ldots, \infty. \tag{1.5}$$

The *p*-value for the same hypothesis testing is $p_2 = P(y \geq 12 \mid \theta = 0.5) = 1 - [P(y = 11 \mid \theta = 0.5) + \ldots + P(y = 3 \mid \theta = 0.5)] = 0.0327$, meaning that the H_0 is rejected at significance level $\alpha = 5\%$. As the likelihoods in the two models are proportional (the likelihood in the first experiment is $f(x \mid \theta) = \binom{12}{3}\theta^3(1-\theta)^9$ while is $f(y \mid \theta) = \binom{11}{2}\theta^3(1-\theta)^9$ in the second), the inference should be the same. However, the inference is different, violating the likelihood principle. In contrast, in the Bayesian framework the posterior distribution is the same in the two experiments, since the proportionality constant cancels in the calculation of the posterior.

Another advantage of the Bayesian framework is that it deals in a natural way with nuisance parameters through marginalization (equation 1.3). If marginalisation is not possible, Markov chain Monte Carlo algorithms can be used to simulate from the posterior and keep only the samples from the parameters of interest. An alternative approach is to calculate the posterior distribution of the parameters of interest by replacing the other parameters with their maximum likelihood estimates. The technique is called *empirical Bayes*, and is not fully Bayesian. In contrast, in classical statistics one has either to estimate all parameters involved in the model or use variations of the classical likelihood approach (e.g. profile likelihood) which increase complexity, to estimate only the parameters of interest.

1.3 The impact of the prior

The use of the prior distribution is at the heart of the Bayesian inference and is either the primary advantage over the classical approach or the major disadvantage. Sometimes prior information for a parameter of interest might be available before collecting any data and this should be used for statistical inference. For example, if one is interested in estimating the divergence time of two species using molecular data, some prior information may be available from the fossil record; the age of the oldest fossil belonging to one of the two species could serve as a minimum bound for their divergence time. The Bayesian approach provides a straightforward way to incorporate any such information through the prior.

In some cases researchers might have different prior beliefs about a parameter due to disparate past observations or due to incongruencies in the current knowledge. Moreover, the representation of the same information by a statistical distribution could not be unique as different priors might seem equally reasonable. As a result, researchers may use different priors which may result in differences in the posterior inference. The problem could be more

severe when there is no prior information around the parameter of interest. In such a case the prior should represent total ignorance. However, with no information about the parameter it is quite unclear which prior is more reasonable. An approach could be to use a uniform distribution over the range of the parameter where any value is equally likely, but this prior might lead to contradictions as it is non invariable to non-linear transforms. One could use priors invariable to reparameterizations, called "Jeffrey's priors", but these may sometimes be improper (do not integrate to 1). For example, the Jeffrey's prior for the mean of a normal distribution (μ) with known variance is $f(\mu) \propto 1$, which does not integrate to 1. The use of improper priors may lead to improper posteriors but their use is broadly accepted as long as the posterior is proper. In general, it is very difficult to represent total ignorance and the subjectivity around the specification of the prior has raised major criticisms for the Bayesian approach.

When we analyze large data sets which are typically informative about the parameter of interest and a diffuse prior is used, the prior impact is practically negligible. In fact, with vague priors the likelihood and the Bayesian estimates are both close to the true parameter value and the confidence and credibility intervals are very similar. For example, let's assume that we are interested in the parameter $\theta \in (-\infty, +\infty)$ and we specify the uniform prior $f(\theta) = 1/(b-a)$, $a < \theta < b$. Then, the posterior distribution is $f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{\int_a^b f(x \mid \theta) f(\theta) \mathrm{d}\theta} =$

$= \frac{f(x \mid \theta)}{\int_a^b f(x \mid \theta) \mathrm{d}\theta}$. The constant terms a and b cancel in the calculation of the posterior. If the likelihood lies within the prior interval (a, b) then the value of the integral in the denominator, and thus the posterior, will be the same irrespective of the values a and b. However, if the prior interval (a, b) is misspecified (does not include the true value) the likelihood might be outside the interval and then the posterior will depend on the precise values of a and b. In general, the less informative the prior is (the further apart a and b are) the more likely is to include the likelihood and thus the lower impact is going to exert in the posterior. The posterior, however, can be sensitive to the prior if the data are uninformative about the parameter of interest or if a model with highly correlated parameters is used (see §2.3.5 for such an example).

Prior specification is an important issue in any Bayesian analysis and there is no easy way on how to best elicit prior information. In general, the prior should summarize one's prior beliefs into a statistical distribution and in case there is no prior information it is always prudent to use diffuse priors. In practice, it is always important to examine the impact of the prior through a Bayesian robustness analysis. Alternative priors can be used and any

significant changes in the Bayesian inference should be reported. In any case, if the likelihood dominates the posterior the choice of the prior becomes unimportant in Bayesian parameter estimation.

1.4 Markov chain Monte Carlo techniques

When modeling real-world problems it is usually necessary to build parameter-rich models. Bayesian inference, except for very simple and usually unrealistic models, requires the calculation of high-dimensional integrals, which is not always practical to compute. Powerful simulation algorithms known as Markov chain Monte Carlo (MCMC) have been developed to deal with the issue. Although they were firstly proposed around 1950s (Metropolis, et al. 1953), it was only after the early 90s that they met an extensive use, as the field was revolutionized by the advance of computational resources (Robert and Casella 2011). These algorithms avoid the calculation of integrals and provide a sample from the posterior distribution via a simulation process. As typically large samples are generated from the posterior the algorithms are computationally expensive. However, those techniques are primarily responsible for the great upsurge in popularity of Bayesian statistics since they enable inference in complex real-world applications.

MCMC methods simulate a Markov chain whose stationary distribution is the posterior distribution of the parameters of interest. Let's assume that we are interested in the posterior distribution $f(\theta_1,\ldots,\theta_p \mid X)$. Then a sample from the posterior can be obtained by the Metropolis-Hastings (MH) algorithm (Metropolis, et al. 1953; Hastings 1970). A sketch of the algorithm is as follows:

1. Set initial state $\boldsymbol{\theta}^j = (\theta_1^j, ..., \theta_p^j)$, $j = 0$.
2. In the j+1 iteration propose a new state $\boldsymbol{\theta}^* = (\theta_1^*, ..., \theta_p^*)$ drawn from a proposal density $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^j)$.
3. Set the next value $\boldsymbol{\theta}^{j+1}$ in the chain $\boldsymbol{\theta}^{j+1} = \begin{cases} \boldsymbol{\theta}^*, \text{ with probability } \alpha \\ \boldsymbol{\theta}^j, \text{ with probability } 1-\alpha \end{cases}$, where $\alpha = \min\left\{1, \frac{f(X \mid \boldsymbol{\theta}^*)}{f(X \mid \boldsymbol{\theta}^j)} \frac{f(\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}^j)} \frac{q(\boldsymbol{\theta}^j \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^j)}\right\}$.
4. Increment j. Go back to step 2 until $j = J$, where J is a predetermined number of iterations.

Note that the algorithm does not involve calculation of the marginal likelihood as it cancels out in the calculation of the acceptance probability α. Moreover, in steps 2 and 3 it is not

necessary to update all parameters at once. It is actually advisable to update each parameter separately or create groups of parameters (i.e. grouping correlated parameters) and update the groups one by one. Updating all parameters together is complicated and might lead to poor performance of the algorithm (i.e. very low acceptance rates and thus poor efficiency). The states $(\theta_1^1, ..., \theta_p^1), ..., (\theta_1^J, ..., \theta_p^J)$ are a sample from the joint posterior $f(\theta_1, \ldots, \theta_p \mid X)$. Note also that the ith ($i = 1, ..., p$) component of each state is a sample from the marginal posterior distribution $f(\theta_i \mid X)$. We can thus summarize the posterior information for any of the parameters. For example, for the parameter θ_i we can approximately estimate the posterior mean through $\mathrm{E}[\theta_i \mid X] \approx \frac{1}{J}\sum_{j=1}^{J}\theta_i^j$, or the variance and the 95% HPD interval.

There are some important remarks that should be mentioned and that one should bear in mind in every implementation of the MH algorithm. First, the algorithm does not sample from the posterior distribution from the first iteration and usually several iterations are required until the algorithm converges to the posterior distribution. If the proposal density specifies an irreducible and aperiodic chain, meaning that the chain is able to visit all the parameter space with no period, then the convergence is guaranteed. These conditions are easily met for the vast majority of chains that we can construct. Thus in every implementation of an MCMC algorithm some first iterations are considered as a burn-in period and are discarded from any subsequent analysis. In all MCMC applications it is extremely important to check for convergence. Several diagnostic tools and tests have been proposed (e.g. the estimated potential scale reduction by Gelman and Rubin 1992), but none of the approaches can guarantee convergence (however, all tests can reject convergence). A common approach is to run the MCMC several times from different starting values and check that the estimates (i.e. posterior means) are the same in all runs. If convergence is not achieved a longer burn-in might fix the problem.

All MCMC algorithms create a dependent sample from the posterior. So, it is common to *thin* the chain by keeping states at specific number of iterations, as the thinned sample has a reduced autocorrelation. The major advantage is that we save computer disk space and avoid computationally intensive calculations when producing summary statistics.

The performance of the MH algorithm highly depends on the proposal density. Generally, there are no guidelines for the choice of the proposal density and a lot of effort has been put by statisticians for the development of efficient proposals (see e.g. Yang and Rodriguez 2013; Yang 2014, chapter 7). For example, one may use as a proposal density the normal distribution with mean the current state θ_i and variance v (called step length). Note that usually the proposed value of a parameter θ_i^* depends on the current state θ_i, although this is not a requirement. The choice of the value of v is also important. If v is too small the

proposed values will be very close to the current values and the chain will move with tinny steps whereas if v is too large the proposed values will be markedly different and most proposals will be rejected, resulting in slow convergence. Moreover, in both cases the sampled states are highly correlated and the algorithm does not explore the parameter space efficiently (poor mixing). Typically the step length is chosen by trial and error so that the acceptance probability is ~30%. This empirical rule is followed in all MCMC implementations presented in the following chapters.

1.5 Statistical inference methods and simulations

In this thesis we analyse molecular data from a broad range of animal species to study natural selection and species diversification times towards a better understanding of the underlying evolutionary mechanisms. We built new statistical methods, tested their performance over pre-existing methods and used them to analyse molecular data sets. Simulations played an important role as they helped us to compare statistical inference methods.

Several statistical methods might be available to analyse empirical data sets and study particular problems such as the genetic relationships among different species. Deciding which of these methods to use might be confusing, but simulation studies could be used to provide some guidance (Huelsenbeck 1995a). Since a statistical model is always an approximation of the real world process, simulations may never be fully representative of reality. However, one is nearly entirely free to simulate data under particular assumptions. Then, it is possible to examine the performance of a statistical inference method when (i) all assumptions of the method are met or (ii) when one or more assumptions are violated in certain ways. In this way we can evaluate the reliability and robustness of the inference method. For example, we can test whether a method can recover the true divergence times on a phylogeny when a particular assumption (e.g. gene trees match the species tree) is violated to a certain extent (e.g. see §4). Via simulations we can also evaluate the statistical properties of an inference method such as the bias and variance of parameter estimate, and the type I & type II errors of a statistical test.

Designing a simulation study so that the results are useful and of general applicability requires careful consideration (e.g. see Burton, et al. 2006; Sokolowski and Banks 2010, for some general guidelines). Many independent replicate data sets are simulated under the same model with the same parameters. The simulated data sets should represent data sets collected in real life. For example, one may use parameter estimates from real data sets (of different features) to conduct the simulation. The simulated data are then analysed with the statistical

inference methods of interest and the results are compared against the true parameter values. Also, it is important for the values of the parameters used to simulate data to cover a wide range (e.g. see §3.3.2). If a method performs well over a wide range of values for a parameter used in the simulation (such as branch lengths of 0.01, 0.1, 1 and 10) then it is sensible to expect it to perform well for parameter values not used in the simulation but within the range (such as branch length of 0.05). In case of Bayesian inference methods where the inference for a parameter is its posterior distribution, one may use the median or mean of the posterior as a point estimate to compare against the true parameter value.

Measures of performance of a method could include accuracy, precision and coverage. To measure accuracy one may use the relative error which is defined as the bias over the true parameter value (i.e. relative error = $\frac{\hat{\theta}-\theta}{\theta}$, or $\frac{\hat{\theta}-\theta}{\theta}100\%$ as a percentage). The relative error is used as an indicator of how good an estimate is relative to the value of the parameter being estimated, providing also the direction of the bias (i.e. underestimate or overestimate). To assess precision, a useful measure is the standard error or the confidence interval width, both showing the variability of an estimator. The mean square error (MSE) of an estimator is another useful measure which combines bias and variance; MSE($\hat{\theta}$) = E[$(\hat{\theta}-\theta)^2$] =Var($\hat{\theta}$) + [E($\hat{\theta}$)$-\theta$]^2. Coverage is also important. By coverage we mean the percentage of times that the confidence interval (in a ML framework, or the credibility interval in a Bayesian setting) contains the true parameter value over the simulated replicate data sets. Note that for 95% confidence (or credibility) intervals the coverage is expected to be around the nominal level of 95%. Values much less than 95% are considered troublesome as they lead to higher than expected type I error rate. The power of a statistical method for a given significance level can also be estimated as the percentage of times that the null hypothesis is rejected, when the null hypothesis is indeed false. For example, one may simulate protein-coding genes from two species assuming a true nonsynonymous/synonymous rate ratio of $\omega = 2$, and then analyse the data and calculate the power for rejecting the null hypothesis H_0: $\omega = 1$. The type II error is calculated as 1 − power. Since results across different measures may vary leading to different conclusions, it is advised to always use more than one criterion to evaluate a method. For example, using the relative error only, one may infer superiority of a method with respect to bias but confidence intervals may suggest high uncertainty. This information could be valuable and will be overlooked if someone concentrates only in assessing accuracy. Generally there is a trade-off between bias and variance as some methods might be more accurate but less precise than others. Statistical methods producing accurate but highly variable estimates (i.e. large confidence intervals) or highly precise but biased are of little practical importance.

An important property of a statistical inference method is robustness which is the ability of the method to estimate the true parameter values even in cases where some of its assumptions are violated. Violation of a method's assumptions is quite common in the analysis of real data. There are several ways in which a method's assumptions might be violated. Some of these violations are known and can be tested, but some may be unknown. It is often impossible to examine the effect of all possible ways and extents of violations in an exhaustive manner. Alternatively, robustness to selected model violations is examined at a time. First, the performance of a method should be evaluated when all of its assumptions are met. Then, the performance of the method is evaluated when one particular assumption is violated at a time, then when two assumptions are violated at the same time and so on. The extent of violation can either be informed from real data (e.g. from past real data analyses) or can be high enough to allow the assessment of the limits of a method's robustness. When comparing the robustness of several methods extra care should be given not to violate assumptions of different methods in different ways because this may create conditions favourable to one of the methods, leading to wrong generalizations about the relative robustness of the methods (Huelsenbeck 1995b). For example, when comparing ML to Bayesian phylogenetic reconstruction the same evolutionary model, e.g. Jukes-Cantor (JC69; Jukes and Cantor 1969) should be used in both approaches. Otherwise, one cannot infer whether the superiority of a method is due to the method itself or due to the evolutionary model used. In general, robustness is an important property and is crucial in selecting among methods.

When analysing real data it is always advised to formulate particular hypotheses before observing any data. To test the hypotheses one should not focus only on the point estimates but consider the uncertainty around them as well. For example, we may are interested in testing whether positive selection has been operating in a gene from species A and B. Let's assume that we analyse a molecular alignment from these species and that for the nonsynonymous/synonymous rate ratio we obtain an estimate $\hat{\omega} = 3$. We can't claim that positive selection (inferred when $\omega > 1$) has been operating based solely on the point estimate, because the high ω value could be just a result of chance effects (i.e. random sampling). *P*-values or confidence intervals (or posterior probabilities and credibility intervals in a Bayesian setting) account for uncertainties in the parameter estimates. Thus, in the previous example positive selection is plausible only in case the confidence interval around $\hat{\omega} = 3$ does not include 1. Biological significance is important as well. For example, an estimate $\hat{\omega} = 1.1$ although statistically significant (e.g. due to large sample size) might be of little biological importance. *P*-values could only inform on the statistical significance and thus it is advised to report both *p*-values (or confidence intervals) and the point estimates (Nuzzo 2014). On the other hand, one should always bear in mind that a non-significant

result does not necessarily mean that the null hypothesis is true but that the method may lack power to reject it.

2 Bayesian applications in molecular evolution

2.1 Programs for Bayesian phylogenetic inference

Bayesian techniques were introduced in molecular phylogenetics in the 1990s mainly to estimate phylogenetic trees from sequence alignments (Rannala and Yang 1996; Mau and Newton 1997; Yang and Rannala 1997; Mau, et al. 1999). The early applications were simple as they assumed the strict clock (constant rate of evolution across the branches of a tree) and they used simple nucleotide substitution models. The following years the field met an explosive growth and many Bayesian computer programs are now available to address several important biological problems using more complex and realistic models (Table 2.1). For example, the program MrBayes performs phylogeny reconstruction using complex models of nucleotide, amino-acid and codon substitution and accounts for rate variation among sites (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). It also allows the analysis of heterogeneous data sets consisting of different data types (e.g. nucleotide and protein) in a combined analysis. The most recent release of the program (MrBayes3.2) has several computational advances (e.g. new proposals, improved convergence, faster likelihood calculations) and provides more output options such as ancestral states and ages of internal nodes (Ronquist, Teslenko, et al. 2012). The program BEAST (Bayesian Evolutionary Analysis Sampling Trees) estimates rooted phylogenies and divergence times using relaxed clock and parametric coalescent models and allows analysis of heterochronous (time-stamped) sequence data (Drummond, et al. 2006; Drummond and Rambaut 2007). PhyloBayes accounts for heterogeneity in evolutionary processes among sites using mixture models of amino acid substitution and provides reliable phylogenetic reconstruction of old phylogenies (Lartillot and Philippe 2004; Lartillot, et al. 2007; Lartillot, et al. 2009).

MCMCTREE may be the first Bayesian program for phylogenetic inference (Rannala and Yang 1996; Yang and Rannala 1997). It is part of the PAML package (Yang 2007) and is popular for Bayesian estimation of species divergence times. It estimates the node ages on a fixed species phylogeny using information from molecular data (nucleotide or amino acid alignments) and the fossil record (in the form of time prior). It allows inference using various nucleotide and amino acid substitution models and implements strict and relaxed clock models. It is computationally efficient as it implements an approximate calculation of the likelihood (dos Reis and Yang 2011) and allows the estimation of divergence times using

Table 2.1: A list of some popular Bayesian phylogenetic programs

Program	Description	Link	Reference
BEAST	Inference of rooted, time-measured phylogenies using the clock and relaxed-clock models. Suitable for analysis of nucleotide and amino acid alignments and morphological data. Strong for analysis of time-stamped data.	http://beast.bio.ed.ac.uk/	(Drummond, et al. 2012)
BPP	Inference of ancestral population size, species divergence times and species delimitation under the multi-species coalescent model. Species tree estimation can also be performed.	http://abacus.gene.ucl.ac.uk/software.html	(Yang and Rannala 2010; Yang 2015)
MCMCTREE	Part of the PAML package for species divergence time and rate estimation using multiple fossil calibrations on a fixed rooted phylogeny. Similar to the MULTIDIVTIME program.	http://abacus.gene.ucl.ac.uk/software/paml.html	(Yang 2007)
MrBayes	Phylogenetic analysis from heterogeneous data (nucleotide, amino acid, morphological data) in a combined analysis. Abundance of evolutionary models. High computational efficiency.	http://mrbayes.sourceforge.net/	(Ronquist, Teslenko, et al. 2012)
MULTIDIVTIME	The first Bayesian program for estimating rates of molecular evolution and species divergence times using a relaxed clock model and approximate likelihood calculation.	http://statgen.ncsu.edu/thorne/multidivtime.html	(Thorne, et al. 1998; Thorne and Kishino 2002)
PhyloBayes	Bayesian MCMC program for phylogenetic reconstruction. Implements the infinite mixture model (CAT) to account for heterogeneity in evolutionary processes among sites.	www.phylobayes.org	(Lartillot, et al. 2009; Lartillot, et al. 2013)

multilocus sequence data of many species. MULTIDIVTIME is another useful program to study divergence times and rates on a phylogeny (Thorne, et al. 1998; Kishino, et al. 2001). The program is very similar to the MCMCTREE and it mainly differs in the construction of time and rate priors (Inoue, et al. 2010). It requires an out-group species to root the in-group tree and allows for different substitution models to be used for multiple data partitions in a combined analysis.

Recently, powerful Bayesian methods were developed to estimate species trees under the multispecies coalescent model, accommodating conflicts among gene trees (Liu, et al. 2009). These methods estimate the species tree from a multilocus sequence alignment and account for errors in the estimation of gene trees. BEST (Bayesian Estimation of Species

Trees) estimates the species tree along with the species divergence times and ancestral population sizes using the posterior distribution of gene trees estimated by MrBayes (Liu and Pearl 2007; Liu 2008). *BEAST (read star-BEAST) estimates the species tree as well as the gene trees and has been found to outperform BEST in estimation of divergence times and ancestral population sizes (Heled and Drummond 2010). The BPP program, in its early release, was designed to estimate simultaneously ancestral population sizes and divergence times on a fixed species phylogeny under the multispecies coalescent (Rannala and Yang 2003). Different numbers of sequences are allowed at different loci. The program was later extended allowing for species delimitation inference from a user-specified guide tree (Yang and Rannala 2010; Rannala and Yang 2013). The most recent version of the program (BPP3.1) performs simultaneously Bayesian estimation of species tree and species delimitation (without the need for a user-specified guide tree) under the multispecies coalescent model (Yang and Rannala 2014; Yang 2015).

All these programs use powerful MCMC algorithms to search the parameter space, as calculation of the marginal likelihood is impossible even for data sets with only a few taxa. The programs are efficient and allow Bayesian MCMC inference from several taxa in realistic time frames.

Bayesian methodology is currently applied in a broad range of biological problems such as the evolutionary relationships among species, population demographic histories, timings of speciation events and natural selection. In this thesis we will focus on Bayesian methods to study natural selection and species divergences times. So, in the following sections we describe existing Bayesian techniques to address those problems together with some theory necessary for the better understanding of the subsequent chapters.

2.2 Estimating the mode and strength of natural selection on a protein

2.2.1 Natural selection

Natural selection has always been of particular interest to evolutionary biologists since its introduction by Charles Darwin in his book *On the origin of species* (Darwin 1859). Mutations on a gene may change the amino acid sequence of the encoded protein with potential changes in protein function, which can further affect the fitness of an individual compared to the rest of the population. If the mutation offers a survival or fertility advantage

to the individual (positive selection) it is likely to pass to the progeny and then further spread to the whole population until its fixation (all individuals will carry the mutation). In contrast, if the mutation is deleterious (purifying selection), an individual carrying the mutation would have a survival disadvantage and may not survive long enough to produce progeny. In this case the mutation will eventually get lost. If the mutation is neutral (neither advantageous nor deleterious) then its fate is determined by genetic drift. Genetic drift is the random fluctuation of allele frequencies due to the stochastic nature of the reproduction process. Genetic drift may affect fixation of advantageous or deleterious mutations as well, if those do not have a strong effect in fitness and its contribution is more important in small populations (Hedrick 2011).

Natural selection occurs when individuals carrying a specific genotype (in other words specific mutations) are better adapted to the environment and have a survival and/or fertility advantage. There are different types of natural selection such as directional, stabilizing, diversifying and balancing selection.

In directional selection the mutant allele provides survival and/or fertility advantage and is thus favored over all other alleles leading to an increase in the frequency of the mutant allele. Directional selection occurs usually under environmental changes or after species migrate to a new environment. In such a case, individuals carrying the advantageous allele are able to pass it to more offspring than those they lack it and so eventually the frequency of the advantageous allele in the population increases. A famous example of directional selection is the Industrial Melanism of the peppered moth population in England (Majerus 2008). Before the Industrial Revolution the majority of peppered moths were white, while dark moths were less frequent. Due to industrialization the air became polluted and thus the barks of the trees blackened. The dark-colored moths obtained a fitness advantage over their white counterparts since they could camouflage themselves more efficiently in the dark barks, and thus their frequency increased.

Diversifying selection occurs when extreme genotypes are favored over intermediate genotypes. An example of diversifying selection may concern the colour of lizards living in an environment with only black and white rocks. Let’s assume that lizards have black, white (extreme phenotypes) or grey (intermediate phenotype) skin. Then given that in the lizard’s habitat there are only black and white rocks, the black and white skin offers greater protection from predators and thus the population of grey lizards will decrease over time. The population of lizards experiences diversifying selection for the extreme phenotypes of the skin colour. Diversifying selection is rare but is an important force of evolution as not only maintains polymorphism, but because it favors divergent traits it may cause speciation (Smith 1962; Rice and Salt 1988).

Stabilizing selection is the opposite of diversifying selection and occurs when intermediate genotypes provide a selective advantage over extreme genotypes causing the population to gradually shift towards intermediate variants. Stabilizing selection was thought to be the most common type of natural selection (Charlesworth, et al. 1982), but recent studies have shown that this may not be the case (Kingsolver, et al. 2001). Moreover, it is believed that stabilizing selection reduces genetic variation. However, since it is hard to measure the strength of stabilizing selection, this is based more on intuition rather than on scientific evidence (Barton and Keightley 2002). Many traits such as the number of offspring in a population of a mammal species could be claimed to be under stabilizing selection. Although in these cases intermediate phenotypes have reproductive advantage, it is hard to know whether this is actually attributed to stabilizing selection (Kondrashov and Turelli 1992).

Balancing selection is another form of selection according to which multiple alleles are maintained in a population, usually due to forces favoring heterozygotes (*heterozygote advantage*; Hedrick 2011). Other forces such as frequency-dependent selection, where rare genotypes are advantageous, or due to varying selection in space and time, where different genotypes are advantageous in different environments or time periods may also lead to maintenance of multiple alleles in the gene pool of a population. Thus balancing selection helps to maintain genetic polymorphism. A fairly-known heterozygote advantage example concerns the sickle cell anemia in humans, a hereditary condition which damages the red blood cells (Pauling, et al. 1949). Homozygote individuals for the abnormal allele HgbS of the haemoglobin gene have damaged red blood cells (i.e. rigid and sickle-shaped) which can't carry as much oxygen as the normal red blood cells (flexible and disc-shaped) causing tiredness and breathlessness. Heterozygote individuals carry the normal haemoglobin gene HgbA and the defective one and may suffer from similar problems from time to time. However, the heterozygote carriers are resistant to malaria parasites and thus the heterozygote genotype is advantageous in regions where malaria exists, as the normal homozygotes suffer from malaria and the abnormal heterozygotes suffer from sickle cell anemia (Allison 1954).

To study natural selection, protein-coding regions of DNA offer a great advantage over non-coding regions because one can distinguish synonymous and nonsynonymous mutations. Synonymous mutations are those that do not change the amino acid in the protein encoded by the codon, whereas nonsynonymous mutations do change it. The most popular method to test for positive selection in protein-coding genes is based on the ratio ($\omega = d_N/d_S$) of nonsynonymous (d_N) to synonymous (d_S) rates and assumes that selection is applied to the protein although mutations occur at the DNA level (Miyata, et al. 1979; Miyata and Yasunaga 1980). The d_N is defined as the number of nonsynonymous mutations per

nonsynonymous site; similarly for d_S. The fixation rate of the nonsynonymous mutations relative to that of synonymous reflects the type of selection. If nonsynonymous mutations are deleterious, because of purifying selection their fixation rate will be less than that of synonymous mutations and thus $d_N < d_S$ and $\omega < 1$. If nonsynonymous mutations are advantageous, their fixation rate will be higher than that of synonymous mutations because of positive selection and thus $d_N > d_S$ and $\omega > 1$. If the nonsynonymous mutations are neither advantageous nor deleterious (neutral evolution) selection does not have an effect on fitness and then those are expected to become fixed at the same rate as synonymous mutations, so that $d_N = d_S$ and $\omega = 1$. Values of ω significantly higher than 1 indicate positive selection with higher values indicating stronger selection.

To better understand the relationship between ω and the underlying selection pressures one should have a look at the neutral theory of molecular evolution (Kimura 1968, 1969; King and Jukes 1969; Kimura and Ohta 1971). The strictly neutral theory suggests that the new mutations are either highly deleterious and removed by natural selection, or have no fitness effect (neutral) and their fixation is random, determined by genetic drift (Kimura 1968). Thus, deleterious mutations have only a small contribution to the genetic variation within species and no contribution at all to that among species. Advantageous mutations are assumed to occur very rarely, hence leaving neutral mutations to be the main source of genetic divergence among species.

Assume that a new mutation occurs in a haploid population, with relative fitness $1+s$, while the common wild type allele has fitness 1. Then, the probability that the mutation will eventually become fixed in the population is

$$P = \frac{2s}{1 - e^{-4Ns}} \tag{2.1}$$

where N is the population size (Fisher 1930). If $s > 0$ the mutation is selectively favored and positive selection is operating. In contrast, if $s < 0$ the mutation is selectively disfavored and negative selection is operating. If $s \approx 0$ (neutral mutation) the probability of fixation becomes $P = 1/2N$ and its eventual fixation is random.

Suppose that in a diploid population mutations occur at rate μ per locus per generation (mutation rate) where a fraction f_0 of them are neutral and the rest $1\text{-}f_0$ are highly deleterious. Then the substitution rate per generation, r_0, is the product of the expected number of neutral mutations per generation times the fixation probability of a neutral mutation. The expected number of neutral mutations per generation is the rate at which mutations occur per locus per generation (μ) times the number of alleles ($2N$ in diploid populations) times the proportion of the mutations that are neutral (f_0). Thus the substitution rate is

$$r_0 = \mu \times 2N \times f_0 \times \frac{1}{2N} = \mu f_0 \tag{2.2}$$

An important modification to the strict neutral theory was proposed later by Ohta (1973). According to her *nearly-neutral* model the newly arising non-lethal mutations are not necessarily strictly neutral but are allowed to be slightly deleterious, while positive selection is disallowed. Later a newer modification was proposed allowing for a proportion of new mutations to have positive selection coefficients (Ohta 1992). Let's assume under this model that f_s is the fraction of the mutations that are selected and r_s is their substitution rate. Then the rate r_s would be the product of the expected number of selected mutations per generation times the probability of a selected mutation to become fixed (given by 2.1). Thus

$$r_s = \mu \times 2N \times f_s \times \frac{2s}{1-e^{-4Ns}} = f_s \mu \frac{4Ns}{1-e^{-4Ns}} \qquad (2.3)$$

By comparing the relative rate of substitution of selected mutations (eq. 2.3) to that of neutral mutations (eq. 2.1) we obtain a measure (ω) to study the implications of natural selection:

$$\omega = \frac{r_s}{r_0} = \frac{f_s}{f_0} \frac{4Ns}{1-e^{-4Ns}} \qquad (2.4)$$

Assuming that synonymous mutations are neutral (i.e. $d_S = r_0$) and nonsynonymous to be of any kind ($d_N = r_s$) the above ratio can be interpreted as the expected d_N/d_S ratio. The interpretation of ω can be clearer through some examples. (i) Let's assume that positive selection is not operating. Then the number of synonymous mutations fixed per generation (fixation rate) according to eq. (2.2) is $d_S = \mu$ ($f_0 = 1$ since all synonymous mutations are considered to be neutral). Then for the nonsynonymous mutations we consider that a fraction f_s of them are neutral and fix at rate μ, and the rest 1-f_s are deleterious (since positive selection is not operating) and are not fixed. Thus the overall number of nonsynonymous mutations fixed per generation is $d_N = f_s\mu + (1\text{-}f_s)\ 0 = f_s\mu$, so that $\omega = d_N/d_S = f_s$. If all nonsynonymous mutations are neutral ($f_s = 1$; neutral evolution) then $\omega = 1$, and if a fraction of them are neutral ($f_s < 1$; purifying selection) then $\omega < 1$. Thus a value of $\omega < 1$ is indicative of purifying selection. (ii) Now let's assume that positive selection is operating. The number of synonymous mutations fixed per generation is again $d_S = \mu$. For the nonsynonymous mutations we consider that a fraction 1-f_s are deleterious and are not fixed, and a fraction f_s are non-deleterious. Of the non-deleterious mutations we assume that a fraction θ is advantageous and a fraction 1-θ are neutral. The neutral mutations fix at a rate μ while the advantageous mutations fix with probability $P = \frac{2s}{1-e^{-4Ns}}$, (eq. 2.1). If $Ns \gg 1$ then $P \approx 2s$. Thus overall the number of nonsynonymous mutations fixed per generation is d_N = (1-f_s) 0 + f_s(1-θ)μ + $f_s\theta 2N\mu$ 2s, so that $\omega = d_N/d_S = \frac{f_s(1-\theta)\mu + f_s\theta 2N\mu 2s}{\mu}$ = f_s(1-θ) + $f_s\theta$

$4Ns$. Note that if θ is large enough (in particular $\theta > \frac{1-f_s}{f_s}\frac{1}{4Ns-1}$) then $\omega > 1$. For example, if $f_s = 0.4$, $\theta = 0.25$, $s = 0.01$ and $N = 10^3$, $\omega = 4.3$ and thus positive selection is inferred. However, note that a value of $\omega < 1$ does not mean that positive selection has not been operating, but simply that cannot be detected. For example, if the proportion of positively selected sites is very small, e.g. $\theta = 0.025$ with f_s, s, and N as previously, then $\omega = 0.79$ and positive selection is not inferred.

Although ω is a useful indicator of selective pressure at the sequence level there are a few assumptions in the above definition of ω as well as some limitations in detecting positive selection with this measure. An important point is that the selective coefficient s is a property of a particular mutation, while ω is a property of a particular site or group of sites in nucleotide sequences. Therefore, some assumptions are required in order to infer the kind of selection from the estimated ω value. In particular, it is assumed that all nonsynonymous mutations in a specific site have the same selective effect (the same selection coefficient s) and thus the same ω ratio is used at a site for all possible amino acid changes. This is a rather strong assumption which may not be met in many cases. For example, some amino acids are chemically similar to one another and a nonsynonymous mutation leading to a small chemical change is more likely to allow the protein to remain functional with the possibility of the mutation to be fixed, whereas a large chemical change is more likely to cause protein malfunction leading to loss of the mutation. However, incorporating different ω ratios for different amino acid changes at a site is not straightforward, as the relationship between amino acid changes and the effects of the modified chemical properties is poorly understood (Yang, et al. 1998; Zhang 2000). Moreover, defining positive selection on a model accounting for amino acid chemical properties is unclear (Yang and Bielawski 2000). Another assumption is that mutations at different sites evolve independently of one another (but see Goldstein, et al. 2015). This is the case for interspecific data given that not many strongly selected mutations are occurring at the same time. If this assumption is not met then the selection coefficient inferred from the estimated ω ratio will be an underestimate (Nielsen and Yang 2003). Indeed, dos Reis (2015) using the mutation-selection model of Halpern and Bruno (1998) showed that eq. (2.4) constitutes a reasonable lower bound on ω. A major limitation of the ω ratio is that it can be used to study natural selection only on the protein-coding regions of a genome although this may operate elsewhere (i.e. regulatory regions of a gene) and as long as the region under study does not overlap with another protein-coding region (Monit, et al. 2015). Such overlapping reading frames are found in viral and bacterial genomes. In addition, because the ω ratio detects positive selection only when the rate of nonsynonymous substitutions is higher than the rate of synonymous, natural

selection that does not lead to more nonsynonymous changes, as is the case of balancing selection may not be detected by ω (Yang and Bielawski 2000).

2.2.2 Codon models

Models of codon evolution offer a clear advantage in studying natural selection over the nucleotide and amino acid models as they use simultaneously the nucleotide information in the DNA and the structure of the genetic code, and thus distinguish between nonsynonymous and synonymous mutations. The unit of data is the codon and substitutions among codons at any particular codon site are described by a Markov chain running along a phylogenetic tree. Changes among codon sites are assumed to occur independently. The first codon models were proposed in the mid 90s by Goldman and Yang (1994) and Muse and Gaut (1994) and were found to produce reliable estimates of the nonsynonymous/synonymous rate ratio within the ML framework. They also account for other biologically important measures such as the transition/transversion rate ratio (κ) and the codon frequencies.

Yang and Nielsen (1998) proposed a simplified version of the model of Goldman and Yang (1994), which incorporates explicitly the ω ratio. It is very flexible to search for adaptive evolution and thus has been widely used. According to this model the instantaneous substitution rate from codon i to codon j ($i \neq j$) is given by the matrix $Q = \{q_{ij}\}$, where

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition,} \end{cases} \quad (2.5)$$

with π_j to be the equilibrium frequency of codon j. The different states of the Markov chain are the 61 sense codons ($i, j = 1,...,61$) as stop codons are not considered since they are assumed not to occur within protein-coding genes. The diagonal values q_{ii} are defined so that the sum of each row is zero, $q_{ii} = -\sum_{i \neq j} q_{ij}$. Because time and rate are confounded the Q matrix is usually rescaled so that the average rate is $-\sum_i \pi_i q_{ii} = 1$. The transition probability matrix $P(t) = \{p_{ij}(t)\}$is then given by

$$P(t) = \exp(Qt), \quad (2.6)$$

where $p_{ij}(t)$ is the probability that codon i is replaced by codon j after time t. Since the average substitution rate is 1, the time t is measured by the expected number of nucleotide substitutions per codon (the time t is the distance $d = -t\sum_i \pi_i q_{ii} = t$).

The rate matrix Q specifies an irreducible (the chain can jump from state i to any state j) and aperiodic Markov chain with a unique stationary distribution $\boldsymbol{\pi} = (\pi_1,...,\pi_{61})$. The codon substitution process along any phylogenetic tree is assumed to be stationary, meaning that the codon frequencies π_j are the same throughout the evolutionary time. Moreover, the Markov process is time-reversible since $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$, for $i \neq j$ and for any t. This practically means that the pattern of evolution is the same whether time runs forward or backwards, and the likelihood calculated on a phylogenetic tree will be the same irrespective of the location of the root on the phylogeny. Moreover, the Q matrix is independent of time (homogeneous) so that the transition probability $p_{ij}(t)$ is the same in any part of the tree for the same time interval t. The assumptions of time-reversibility, stationarity and homogeneity are the same as in any nucleotide model and although they might be biologically unrealistic they are mathematically convenient. Some of those assumptions can be easily relaxed (e.g. the homogeneity; Yang and Roberts 1995) and might allow inference on important biological questions which is trivial with the simplified models (e.g. identifying pathogens host shifts; dos Reis, et al. 2009; Tamuri, et al. 2009). However, relaxation of other assumptions (e.g. time reversibility to infer the root of a phylogeny; Barry and Hartigan 1987) can lead to complex calculations with no significant gain in inference (Yang 1994a). The codon model described here will be used in chapter 3 to estimate ω through a Bayesian approach.

2.2.3 Techniques to identify positive selection

2.2.3.1 Likelihood techniques

When the codon model specified by equation (2.5) is applied to a phylogeny it is assumed that the ω ratio is the same across all lineages and across all sites in the alignment. Thus positive selection will be detected only when the average ω across all lineages and sites is higher than 1. This is a stringent criterion and may result in low power because positive selection may act in an episodic manner (Messier and Stewart 1997) and affect only a few sites in the protein (Hughes and Nei 1988; Hughes, et al. 1990). Branch models have been developed to relax the assumption of the constant ω ratio across the lineages of a tree (Yang 1998). Lineages on the tree are a-priori divided into *foreground* and *background* branches.

The foreground branches are those of interest and have a ω ratio (denoted ω_f) different than the background branches (ω_b). One can test whether the two ω ratios are the same with a likelihood ratio test (LRT) between this model and another having the same ω in all branches. Similarly, one can test whether $\omega_f = 1$ by performing a LRT between a model where ω_b, ω_f are free to vary and another where ω_b is free and $\omega_f = 1$.

Random-sites models relax the assumption of constant ω ratio across the sites of a protein by assuming that each codon belongs to a site class with a probability (Nielsen and Yang 1998; Yang, et al. 2000b). For example, the M2a model assumes three site classes: one with $\omega_0 < 1$, another with $\omega_1 = 1$ and a third with $\omega_2 > 1$, with a codon to belong to each site class with probabilities p_0, p_1, $p_2 = 1-p_0-p_1$, respectively (Yang, et al. 2005). One can test for positive selection by comparing this model, which allows some sites to have $\omega > 1$, with the M1a model which does not (it has two site classes with $\omega_0 < 1$ and $\omega_1 = 1$ with probabilities p_0 and $1-p_0$, respectively) using a LRT.

The branch-site models allow the ω ratio to vary across both the branches of the phylogeny and the sites of a protein and can detect positive selection operating in some sites along specific branches, which might be more realistic (Yang and Nielsen 2002). The branches on a phylogeny are *a priori* divided into foreground and background branches, as in branch models, while site classes for ω are also defined as in site models. The branch-site model of Yang et al. (2005) has four site classes: site class 0 includes codons with $0 < \omega_0 < 1$ along all lineages; site class 1 consists of neutrally evolving codons ($\omega_1 = 1$) along all lineages; site class 2a includes codons that are under purifying selection on the background branches but under positive selection ($\omega_2 > 1$) on the foreground branches; and site class 2b includes neutrally evolving codons on the background branches but under positive selection ($\omega_2 > 1$) on the foreground branches. A codon belongs to each site class with probability p_0, p_1, $(1 - p_0 - p_1)p_0/(p_0 + p_1)$ and $(1 - p_0 - p_1)p_1/(p_0 + p_1)$, respectively. The ω_0, ω_1, ω_2 are estimated from the data. A test of positive selection along the foreground branches can be performed by comparing this model with the same but restricting $\omega_2 = 1$ through a LRT. Simulations have shown that this model has better power in detecting positive selection than the branch models. However, it could be conservative (Zhang, et al. 2005).

In LRTs of positive selection based on branch-site and branch models the foreground branches have to be specified *a priori*. If there is no biological information to specify particular branches as foreground the LRT can be applied to several branches but a correction for multiple testing is necessary (Anisimova and Yang 2007). However, applying multiple such tests creates a logical inconsistency as a branch might be specified as a background branch in some tests, although it could have been found to be under positive selection in a previous test. To address the issue Pond et al. (2011) developed a random-effects branch-site model where a site at every branch is assigned to a site class $\omega^- \leq \omega^N \leq 1$

$\leq \omega^+$ with a probability, representing strong, weak purifying selection and positive selection, respectively (Pond, et al. 2010). Then to identify if a branch is under positive selection, the same model is applied by restricting $\omega^+ = 1$ to the branch of interest, and a LRT between the two models is performed. With multiple such tests and correction for multiple hits lineages under positive selection can be identified, without restricting background branches to be under purifying or neutral evolution. The model has been found to have similar performance to that of Yang et al. (2005) when the assumptions of the later hold and have better performance when the assumptions are violated (Pond, et al. 2011).

2.2.3.2 Bayesian techniques

In site and branch-site models when a LRT indicates positive selection there is an interest in identifying particular sites in which adaptive evolution has been operating. Bayesian techniques can help to address the issue. The posterior probability that a site h with data x_h belongs to site category k (with ratio ω_k) is given by

$$P(\omega_k \mid x_h) = \frac{p_k f(x_h \mid \omega_k)}{f(x_h)} = \frac{p_k f(x_h \mid \omega_k)}{\sum_j p_j f(x_h \mid \omega_j)}, \qquad (2.7)$$

with p_k to be the probability that the site h belongs to the k category. The category with the highest posterior probability is the most likely for the site h. One can then identify sites under positive selection as those which belong to the category with $\omega > 1$with a high probability (let's say > 95%). For the branch-site model where there are two categories with $\omega > 1$ in the foreground lineages (classes 2a and 2b) one has to sum the two posterior probabilities. Equation (2.7) requires the knowledge of the model parameters such as the ω ratio of each category, the proportions of sites belonging to each category, the transition/transversion rate ratio, the equilibrium codon frequencies, the branch lengths and the phylogenetic tree. Nielsen and Yang (1998) estimated the posterior probabilities using a naive empirical Bayes (NEB) approach where the model parameters where substituted by their maximum likelihood (ML) estimates. However, this approach does not account for uncertainties in the parameter estimates. This might not be a problem in large data sets where there is enough information for the parameters to be precisely estimated but might lead to unreliable posterior probabilities in case of small data sets with low sequence divergences (Anisimova, et al. 2002; Wong, et al. 2004). For example, consider that under the M2a model the maximum likelihood estimates (MLEs) are $\hat{p}_0 = \hat{p}_1 = 0$, $\hat{p}_2 = 1$ and $\hat{\omega}_2 = 1.5$, then the naive empirical Bayes approach would infer that all sites in the sequence are under positive selection with posterior probability 1.

Yang et al. (2005) developed a method which accommodates uncertainties in the MLEs of the ω ratios for each category. They followed a Bayes empirical Bayes approach assigning a prior on the ω ratios and on the proportions and averaged over the priors to obtain the posterior probabilities. Numerical integration was used to calculate the integrals involved. The other model parameters remained fixed to their MLEs as they are considered to be less important for the calculation of posterior probabilities. For example, the tree topology has been found to have only minimal impact (Yang, et al. 2000b; Swanson, et al. 2001). The BEB method applies to both the site and branch-sites models and gives similar estimates to NEB in large data sets. In small data sets the BEB has lower false-positive rate than the NEB but appears to be conservative when a cut-off of 95% is used in the posterior probability (Yang, et al. 2005; Zhang, et al. 2005). In general, it is more difficult to identify sites under positive selection than to test whether such sites exist. The later can be tested with a LRT combining information from all sites in the alignment and thus if many such sites exist the test could be significant. In contrast, information at a single site may not be strong enough (e.g a few substitutions at a site in a few lineages at the tree) to give high posterior probability.

Huelsenbeck and Dyer (2004) developed a fully Bayesian approach to accommodate for sampling errors in all parameter estimates including the tree topology. The high dimensional integrations are intractable analytically and thus they used MCMC to approximate the posterior probabilities. This approach might return more reliable estimates than the BEB in small uninformative data sets (Scheffler and Seoighe 2005), however, because of the iterative MCMC algorithm the method is computationally expensive and slow and is not practical for large data sets. Furthermore, the method implements only the M3 discrete site model (three categories of ω ratio: ω_0, ω_1, ω_2 in proportions p_0, p_1, $1 - p_0 - p_1$).

The site and branch-site models mentioned so far assume that the number of categories is known and *a priori* defined. Huelsenbeck et al. (2006) implemented a more flexible way to account for variation in ω ratio among sites. In their model the number of categories and the ω value for each category are considered random variables and are estimated from the data using a Dirichlet process to assign priors on them. The number of categories is free to vary between 1 and L, where L is the number of codons in the alignment, meaning that each codon is allowed to have its own category. All the parameters of the model are estimated within a Bayesian MCMC framework and inferences of positive selection account for uncertainties in model parameter estimates, branch lengths and topology. The model involves too many parameters and posterior probabilities for inferring sites under positive selection might be affected by the prior on the number of site categories. In analyses of empirical data sets the method gave similar results to that of Yang et al. (2000b).

2.3 Bayesian estimation of species divergence times

2.3.1 The molecular clock

The hypothesis of the molecular clock states that DNA or protein sequences evolve at a constant rate over time which is the same in all species. This hypothesis was proposed by Zuckerkandl and Pauling (1965) based on previous empirical observations that the numbers of amino acid differences in proteins from different species were proportional to the species divergence times estimated from the fossil record (Zuckerkandl and Pauling 1962; Margoliash 1963; Doolittle and Blomback 1964). The amino acid changes that have been accumulated among species were considered to have no or little effect on the structure and function of the protein and thus on fitness. This notion was later formulated by the development of neutral theory (Kimura 1968; King and Jukes 1969). The theory predicts that the rate of fixation of neutral mutations (the substitution rate) equals the neutral mutation rate which is the total mutation rate per generation times the proportion of the mutations that are neutral. If the neutral mutation rate is similar among species then a constant substitution rate across the evolutionary tree of life is possible. Thus the constancy of the substitution rate can be explained by the neutral theory.

The molecular clock was quickly recognised as a valuable tool in the study of molecular evolution. A direct implication of the clock is that the genetic divergence of any two species is proportional to their divergence time. Thus, if the divergence time of two species on a phylogeny is known, say from the fossil record or from a geological event (e.g. island or mountain formation which divides a population in two parts and initiates speciation), one can obtain an estimate of the evolutionary rate from their genetic divergence. Then based on the assumption of rate constancy one can infer the ages of all nodes in a phylogeny. This could be extremely useful in estimating the divergence times of species which have left limited or no marks in the fossil record.

Nowadays it is generally accepted that the clock does not hold for very diverse species (Langley and Fitch 1974), but might be a good approximation for closely related species. Factors such as generation time, population size, metabolic rate, body size, DNA repair mechanisms which may vary dramatically in distantly related species, have been associated with differences in the molecular evolutionary rate (Bromham and Penny 2003; Bromham 2011; Ho 2014). However, because the molecular clock is a valuable tool to study molecular evolution it has not been abandoned. Instead, alternative clock models have been developed which relax the rate constancy across the tree and are used to analyze data from diverse

species. The new relaxed clock models allow the rate to vary along the branches of a phylogeny according to a statistical model. Two widely used relaxed clock models are the independent-rates model (Drummond, et al. 2006; Rannala and Yang 2007) and the autocorrelated-rates model (Thorne, et al. 1998; Kishino, et al. 2001; Rannala and Yang 2007). In the first, the rates vary among branches around a value according to a statistical distribution such as the log-normal, while in the second the logarithm of the rate drifts according to a Brownian motion process.

With the advance of sequencing technologies and the abundance of molecular sequence data, the molecular clock has been widely used to estimate divergence times for a broad range of species. However, its use has raised serious controversies as typically molecular dating studies tend to produce older divergence times than those suggested by the fossil record. Most of those controversies concern important evolutionary events (Cooper and Fortey 1998). Such an example concerns the origin of early animal forms. The fossil record suggests a massive radiation of Bilateria phyla after the Ediacaran-Cambrian boundary (Budd 2008; Maloof, Porter, et al. 2010), 541 million years ago (Ma), but estimates from molecular dating studies are older, placing their origin during the Ediacaran (635 − 541 Ma) or Cryogenian (850 − 635 Ma) periods (Peterson, et al. 2008; Erwin, et al. 2011; dos Reis, et al. 2015) or even earlier (Wray, et al. 1996; Wang, et al. 1999; Nei, et al. 2001). Another example concerns the radiation of mammals followed the extinction of dinosaurs at the Cretaceous−Paleogene boundary (66 Ma). Molecular studies have produced older dates than those expected from fossils, setting up a debate between evolutionary biologists and palaentologists around the true diversification times of mammals (Meredith, et al. 2011; dos Reis, et al. 2012; O'Leary, et al. 2013; dos Reis, Donoghue, et al. 2014). Molecular estimates of angiosperms diversification times are also much older than those suggested by fossils (Bell, et al. 2010).

Part of those incongruences can be attributed to the incomplete fossil record. Fossilization and preservation of fossils can be suspended by environmental factors such as erosion and humidity. Moreover, living organisms with only soft parts are highly unlikely to be preserved. Those factors can lead to systematic preservation biases and produce an imperfect fossil record with a diminishing quality as one goes back in time (Raup 1972). Moreover, there is an important limitation in estimating the age of a clade in a phylogenetic tree solely by the fossil record. The oldest fossil from that clade will always be younger than the origin of the clade either by a few thousand years (which is negligible when dating very ancient events) or by many millions (Benton, et al. 2009). The fossils constitute minimum bounds for the node ages of a phylogenetic tree and since the molecular clock studies attempt to estimate the clade ages any discrepancies become less acute.

Another part of the discrepancies may come from the molecular studies themselves. The early dating studies suffered from methodological deficiencies and limited molecular data. The methods used were too simplistic and failed to account for important uncertainties in the analysis. For example, some used fossil calibrations with fixed ages failing to account for uncertainties in the fossil record, or used the strict clock even for distantly related species (Kumar and Hedges 1998; Peterson, et al. 2004; Peterson and Butterfield 2005). Taxa were also removed to diminish the among lineages rate variation making inefficient use of the data (Peterson, et al. 2004). Moreover, some of the data sets were comprised of just a few genes.

Recently, sophisticated MCMC algorithms have been developed and are capable of analysing large multilocus sequence data to estimate species divergence times. The Bayesian framework provides a straightforward way to accommodate uncertainties in fossils while relaxed clock models and data partitioning are used to deal with rate heterogeneity across lineages and sites. Thorne et al. (1998) and Kishino et al. (2001) developed a Bayesian MCMC algorithm to estimate species divergence times and rates on a fixed phylogeny using the autocorrelated-rates model to describe rate variation across lineages. A similar algorithm was developed later by Yang and Rannala (2006) and Rannala and Yang (2007) which accounts more elegantly for uncertainties in fossil ages and allows inference from multilocus sequence data. In the next sections we pinpoint the most important features of the latter method as this will be used in chapters 4-6 to estimate divergence times from simulated data and animal species.

2.3.2 General framework and calculation of the likelihood

Assume that we have a fixed rooted phylogeny of s species. We denote with D the sequence data, $\mathbf{t}$ the $s-1$ node ages on the phylogeny and $\mathbf{r}$ either rates on the branches as in Rannala and Yang (2007) or on the nodes as in Kishino et al. (2001). We let θ denote the parameters in the substitution model. Write $f(\mathbf{r}|\mathbf{t},\theta)$ and $f(\mathbf{t}|\theta)$ the priors for rates and times, respectively, and $f(\theta)$ for the prior on θ. Then according to the Bayes theorem (equation1.2) the joint posterior distribution of $\mathbf{t}$, $\mathbf{r}$, and θ is given by

$$f(\mathbf{t}, \mathbf{r}, \theta \mid D) = \frac{f(D \mid \mathbf{t}, \mathbf{r}, \theta) f(\mathbf{r} \mid \mathbf{t}, \theta) f(\mathbf{t} \mid \theta) f(\theta)}{f(D)} \tag{2.8}$$

The marginal likelihood $f(D)$ involves integration over $\mathbf{t}$, $\mathbf{r}$, θ and cannot be calculated. An MCMC algorithm is used instead to sample from the joint posterior. The marginal posterior of times $f(\mathbf{t} \mid D) = \int\int f(\mathbf{t}, \mathbf{r}, \theta \mid D) \mathrm{d}\mathbf{r}\mathrm{d}\theta$ can be calculated from the MCMC sample; similarly the marginal posterior of rates and model parameters. The calculation of the

likelihood $f(D|\mathbf{t},\mathbf{r},\theta)$ is straightforward for any substitution model but is computationally expensive for large data sets. Dos Reis and Yang (2011) extended the work of Thorne et al. (1998) and Kishino et al. (2001) and calculated the likelihood approximately by applying the Taylor expansion to the log-likelihood. The method calculates the gradient and the Hessian matrix of the likelihood using the MLEs of the branch lengths and the model parameters before the MCMC run. Then a transformation (e.g. square root, arcsin) is applied to offer better approximation for values away from the MLEs. The approximation is efficient and allows the analysis of large datasets in realistic times (dos Reis, et al. 2012; Jarvis, et al. 2014).

2.3.3 Priors on node ages

Calibrations are of particular importance for species divergence time estimation since it is not possible to obtain time estimates based solely on molecular data. In absence of reliable information about species evolutionary rates, the fossil record can provide an invaluable source of information concerning the node ages on a phylogeny. Fossils are typically used to constrain the node ages between minimum and maximum values. They usually provide good minimum bounds (see §2.3.1) but the specification of maximum bounds is much more complicated. One could use fossils which lack major characteristics of species belonging to the clade of interest from an older geological formation, to set up a maximum constraint (Benton, et al. 2009). Biogeographic events, such as island formations can also serve as plausible maximum bounds if treated with caution (Heads 2005; Goswami and Upchurch 2010).

Using fossil information to calibrate the molecular clock and date species divergences using molecular data is an arduous task. Fossil preservation biases, incorrect placement of fossils on a phylogeny or uncertainties in fossil age estimation may result in erroneous calibrations and thus in biased molecular time estimates (Magallon 2004; Ho and Phillips 2009). However, even when those factors are known, fossils cannot provide point calibrations but instead involve uncertainty which is expressed in the form of a parametric statistical distribution. Yang and Rannala (2006) and Inoue et al. (2010) provided some useful advice for the construction of such calibration densities. For example, the information from the fossil record around the age (t) of a node can be represented with a uniform distribution, $t \sim U(t_L, t_U)$, in case both minimum (t_L) and maximum (t_U) bounds are available for that node. When only minimum or maximum bounds are available the specification of such calibration densities is less trivial (Inoue, et al. 2010) and may involve the use of improper densities (Yang and Rannala 2006). The choice of the calibration densities is

important as those exert a significant effect on molecular time estimates (Inoue, et al. 2010; Warnock, et al. 2012; Magallon, et al. 2013a; dos Reis, et al. 2015). Some attempts have been made to evaluate the quality of fossil-based calibrations (Warnock, et al. 2015), however, a robustness analysis is always useful.

The bounds mentioned in the example above are "hard" meaning that there is zero probability for the node age to be outside the interval (t_L, t_U). This is a strong assumption and might lead to biases if fossil evidence has been misinterpreted. Consequently, maximum bounds are chosen in a conservative manner with potential impact on time estimation. Yang and Rannala (2006) proposed the use of "soft" bounds which assign positive probabilities to all age values. They are constructed by adding a diminishing probability of power or exponential decay beyond a bound that the age of the node is outside the bound. For example, in the uniform calibration density example described above one may assign left and right tail probabilities of 2.5% that the age of the node is outside the bounds. The advantage of soft bounds is that they allow the signal from the molecular data to correct for errors in fossil calibrations or for conflicts among calibrations leading to a more reliable evaluation of the precision of time estimates (Yang and Rannala 2006).

In the Bayesian estimation of species divergence times the calibration densities are used to construct the prior $f(\mathbf{t}|\theta)$ (or $f(\mathbf{t})$) of the node ages on the phylogeny. Yang and Rannala (2006) developed an algorithm to construct the prior $f(\mathbf{t})$ based on the calibration densities and the birth-death process (Kendall 1948). The birth-death process is a mathematical model which describes the dynamical process of speciation and extinction given the birth (λ) and death (μ) rates of a lineage and a species sampling probability (ρ). If $\mathbf{t}_C$ are the ages of the calibrated nodes and $\mathbf{t}_{-C}$ the ages of the non-calibrated nodes then the prior is given by

$$f(\mathbf{t}) = f_{BD}(\mathbf{t}_{-C} \mid \mathbf{t}_C) f(\mathbf{t}_C), \tag{2.9}$$

where $f(\mathbf{t}_C)$ is the joint density of the calibrated nodes and $f_{BD}(\mathbf{t}_{-C}| \mathbf{t}_C)$ is the joint distribution of the uncalibrated nodes, specified by the birth-death process conditioned on the ages $\mathbf{t}_C$. Because of the requirement that a descendant node must be younger than its ancestor, the user-specified densities are truncated by the program to satisfy this condition. This may result in marginal priors (called *effective* priors) that are different than the user-specified calibration densities (Inoue, et al. 2010; Duchene, et al. 2014). Thus one should always run the MCMC chain without data to generate the marginal priors and check that they are reasonable. This prior specification has been implemented in the MCMCTREE program (Yang 2007).

2.3.4 Rate drift models and prior on rates

Bayesian estimation of species divergence times requires a model for the rate drift along lineages and a prior on the evolutionary rates. Currently there are two widely used relaxed clock models to deal with the among-branch rate variation. The relaxed clock model is incorporated in the prior $f(\mathbf{r}|\mathbf{t},\theta)$. For a given locus, the rate (r) at a node given the rate at the ancestral node (r_A) follows a log-normal distribution with density

$$f(r \mid r_A) = \frac{1}{r\sqrt{2\pi tv}} \exp\left\{ -\frac{1}{2tv}\left[\log\left(\frac{r}{r_A}\right) + \frac{vt}{2} \right]^2 \right\}, \; 0 < r < \infty, \tag{2.10}$$

where t is the time duration separating the two nodes. This is equivalent to say that the log rate ($\log r$) follows a normal distribution with mean $\log r_A - vt/2$ and variance vt. Note the mean of the log-normal density is $E(r) = r_A$ and thus the rate at the node is a value around the rate of the ancestral node. The parameter v denotes the violation of the clock with high values meaning serious violation. The joint prior of all node rates $\mathbf{r}$ on the tree is the product of the log-normal distributions from all nodes. This model was proposed by Thorne et al. (1998) and Kishino et al. (2001) and is known as the autocorrelated-rates model. In the implementation of the same model by Rannala and Yang (2007) the log-normal distribution applies to the rates at the midpoint of the branches.

An alternative model was proposed by Drummond et al. (2006) and Rannala and Yang (2007) known as the independent-rates model. For a given locus, the rate at any branch follows a log-normal distribution with density

$$f(r \mid \mu, \sigma^2) = \frac{1}{r\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}\left[\log\left(\frac{r}{\mu}\right) + \frac{\sigma^2}{2} \right]^2 \right\}, \; 0 < r < \infty, \tag{2.11}$$

where μ is the mean rate for the locus and σ^2 is the variance in the log scale. σ^2 measures the departure from the clock with high values (e.g. 0.2) indicating serious clock violation.

For an alignment of L loci the parameters μ_i, $i = 1,...,L$, of the log-normal distribution for each locus can be assigned a gammaDirichlet prior (dos Reis, Zhu, et al. 2014). In the autocorrelated-rates model the μ_i is the rate at the root of the tree in the i locus which evolves according to equation (2.10). A gamma prior (with fixed hyperparameters) is assigned to the mean rate $\bar{\mu} = \frac{1}{L}\sum_{i=1}^{L} \mu_i$ and the total rate $L\bar{\mu}$ from all loci is partitioned across the μ_i rates according to a Dirichlet distribution with parameter α. A higher α means less rate variation among loci. The way and extent of rate variation across lineages is considered independent among loci, although this might not always be realistic. A gammaDirichlet prior can also be assigned to the parameters σ_i^2 or v_i. The gammaDirichlet prior has been found to exert less

influence in the posterior than assuming independent priors among loci (dos Reis, Zhu, et al. 2014).

In the independent-rates model the variance of the rate does not depend on the time and thus the rate can undergo large shifts (depending on the value of σ^2) even for adjacent branches. In contrast, in the autocorrelated-rates model the variance depends on the time and thus the model penalizes large rate variation over short time intervals but allows rate to vary significantly among distant clades. However, the variance increases linearly with the time and in analyses of deep phylogenies this might lead to unduly high rate shifts. Thus the autocorrelated-rates might be more suitable for the analysis of closely-related species while the independent-rates for divergent species. In any case it is always useful to test the robustness of time estimates to the clock model used.

2.3.5 The limits of molecular clock dating

Dating species divergences using molecular data is an unconventional statistical problem. Molecular sequences provide information only about the distances (the product of rates and times) on a phylogeny but not about the times and rates explicitly. Suppose, for example, that two genes sampled from two species are separated by a molecular distance d = 1, meaning 1 nucleotide change per site, on average. We assume that the genes diverged at the same time as the species. Relying only on this information it is impossible to tell whether the species diverged from each other at a rate of 10^{-8} substitutions per site per year (s/s/y) over a period of 50 million years (My), or at a rate of 0.5×10^{-8} s/s/y over a period of 100 My. In fact there are many different combinations of rate and time which might be plausible for the species A and B. Thus to identify the correct combination external information about the time or rate is necessary. Below we extend the example for better understanding.

Suppose that the nucleotide alignment of the two genes consists of n sites with x differences. We assume that the true divergence time is t = 0.5 and the true rate is r = 1. The same rate of evolution (strict clock) is assumed for the two lineages. The time unit is 100 My and thus the true divergence time is 50 million years ago (Ma) and the true rate is 10^{-8} s/s/y, so that their evolutionary distance is $d = 2tr = 1$. The likelihood of the alignment using the Jukes and Cantor (1969) nucleotide substitution model (JC69) is

$$L(D\,|\,d) = p^x(1-p)^{n-x} = \left(\frac{3}{4} - \frac{3}{4}\mathrm{e}^{-4d/3}\right)^x \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4d/3}\right)^{n-x}, \qquad (2.12)$$

where D is the sequence data and p is the expected proportion of different sites in the alignment (p = 0.552 since d = 1). The MLE of the distance is given by

$$\hat{d} = -\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}\right), \tag{2.13}$$

with $\hat{p} = x/n$ to be the observed nucleotide differences in the alignment. Assume now that the alignment is n = 100 sites long and x = 55 differences are observed. We are interested in estimating the divergence time and rate from the molecular data. The model thus contains two parameters (t, r) and the likelihood is

$$L(D\,|\,r,t) = \left(\frac{3}{4}-\frac{3}{4}\mathrm{e}^{-8rt/3}\right)^{x}\left(\frac{1}{4}+\frac{3}{4}\mathrm{e}^{-8rt/3}\right)^{n-x}. \tag{2.14}$$

The likelihood is maximized along the line $r = \frac{\hat{d}}{2t}$, where $\hat{d}$ = 0.991 from equation (2.13) (Figure 2.1A2). Although a MLE is available for d, neither t nor r has a unique MLE. If external information is available for any of the two parameters, an estimate can be obtained using Bayesian methodology. We assume that information from the fossil record suggests minimum and maximum bounds of 40 Ma and 60 Ma respectively, for the divergence time of the species. We represent this information with a gamma prior $t \sim G(100, 200)$, with mean 0.5, meaning 50 My and 95% interval (0.4, 0.6). For the rate we use a diffuse prior $r \sim G(2, 2)$, with mean 1, meaning 10^{-8} s/s/y. The joint prior $f(r, t)$ is the product of the two priors and is shown in Figure 2.1A1. Then, the joint posterior distribution of r and t is

$f(r,t\,|\,D) = \frac{1}{C}L(D\,|\,r,t)\,f(r,t)$, where $C = \int_0^\infty\int_0^\infty L(D\,|\,r,t)\,f(r,t)\mathrm{d}r\mathrm{d}t$ is the normalizing constant. The joint posterior has a mode (Figure 2.1C1) and the means of the marginal posteriors can be used as point estimates for r and t. For example, the posterior mean of t is given by $\tilde{t} = E(t\,|\,D) = \frac{1}{C}\int_0^\infty\int_0^\infty tL(D\,|\,r,t)\,f(r,t)\mathrm{d}r\mathrm{d}t$ and is 0.50, meaning 50 Ma. Similarly, $\tilde{r}$ = 1.03.

We now assume longer genes with n = 1000 sites and x = 552 differences. In that case the distance is more reliably estimated ($\hat{d}$ = 0.999) owing to the larger data set. Note that the likelihood is more concentrated around the line $r = \frac{\hat{d}}{2t}$ in Figure 2.1B2. Using the same prior we get the same inference for the time and rate ($\tilde{t}$ = 0.50, $\tilde{r}$ = 1.01). The joint posterior is more concentrated than in the case of the shorter alignment (Figure 2.1B3) and thus the larger data set led to increased precision of posterior estimates. However, inclusion of more sites in the alignment leads to only a slight reduction in the uncertainty of posterior estimates. For example, the marginal posterior of the rate for n = 5,000 sites is very similar to that for n = 1,000 (Figure 2.2B).

In a typical Bayesian estimation problem as the data increase the prior becomes unimportant. However, this is not the case with the Bayesian estimation of species divergence times because of the confounding nature of time and rate. Increasing the amount of molecular data allows estimation of distances on the phylogeny virtually without error but there is no guarantee that the absolute times and rates will converge to their true values. In fact, even with infinite molecular data the limiting posterior will remain sensitive to the prior. For example, in the two species case described above, consider that we use an incorrect prior $t \sim G(100, 100)$, with mean 1, meaning 100 My and the same prior for the rate to analyze the long alignment (Figure 2.1C1). The joint posterior is now different (Figure 2.1C3) and the posterior estimates are $\tilde{t} = 0.99$ and $\tilde{r} = 0.51$, very far from their true values $t = 0.5$ and $r = 1$ (Figure 2.2C, D). Here, the time prior is very informative and exerts a strong influence in the posterior. If the rate prior is also informative or if both priors are uninformative these will compete each other with an unpredicted effect in the posterior.

Yang and Rannala (2006) and Rannala and Yang (2007) studied the case of an infinite amount of molecular data. According to their "infinite-sites" theory even with an infinite amount of molecular data the posterior time estimates do not converge on point values but to a limiting distribution and thus some uncertainty will always remain. When the amount of molecular data approaches infinity (infinite sites and loci) the 95% credibility interval widths and the posterior time means will fall on a straight line. Thus, in analysis of empirical data such "infinite-sites" plots can be used to evaluate whether inclusion of more molecular data can increase the precision of posterior time estimates (an example concerning the divergence times of 54 metazoan species is shown later in Figure 6.4B).

In general, there are three sources of uncertainty affecting the estimation of species divergence times. The first is the phylogenetic uncertainty (inaccuracies in branch length estimation), caused by finite sequence data. This can be reduced by sampling more molecular data. Dos Reis and Yang (2013a) examined the case of analyzing large but finite sequence data under the clock while Zhu et al. (2015) used relaxed clock models and large multilocus sequence data. The authors developed the "finite-sites theory" according to which a part of the posterior variance of time estimates is due to limited number of sites at a locus and another is due to limited number of loci. To reduce uncertainties of posterior estimates increasing the number of loci seems to be more important than increasing the number of sites within each locus (Zhu, et al. 2015). The second source is uncertainty due to among-branches rate variation which can be improved by sampling more loci. If we assume that all genes in an alignment have the same divergence times (that of the species) but differ in the pattern of evolutionary rate drift, then a long branch in a locus is more likely to be due to an accelerated rate of evolution if the branch is short in all other loci. Thus the use of multiple gene loci seems to be advantageous. The third source is uncertainty in the fossil calibrations

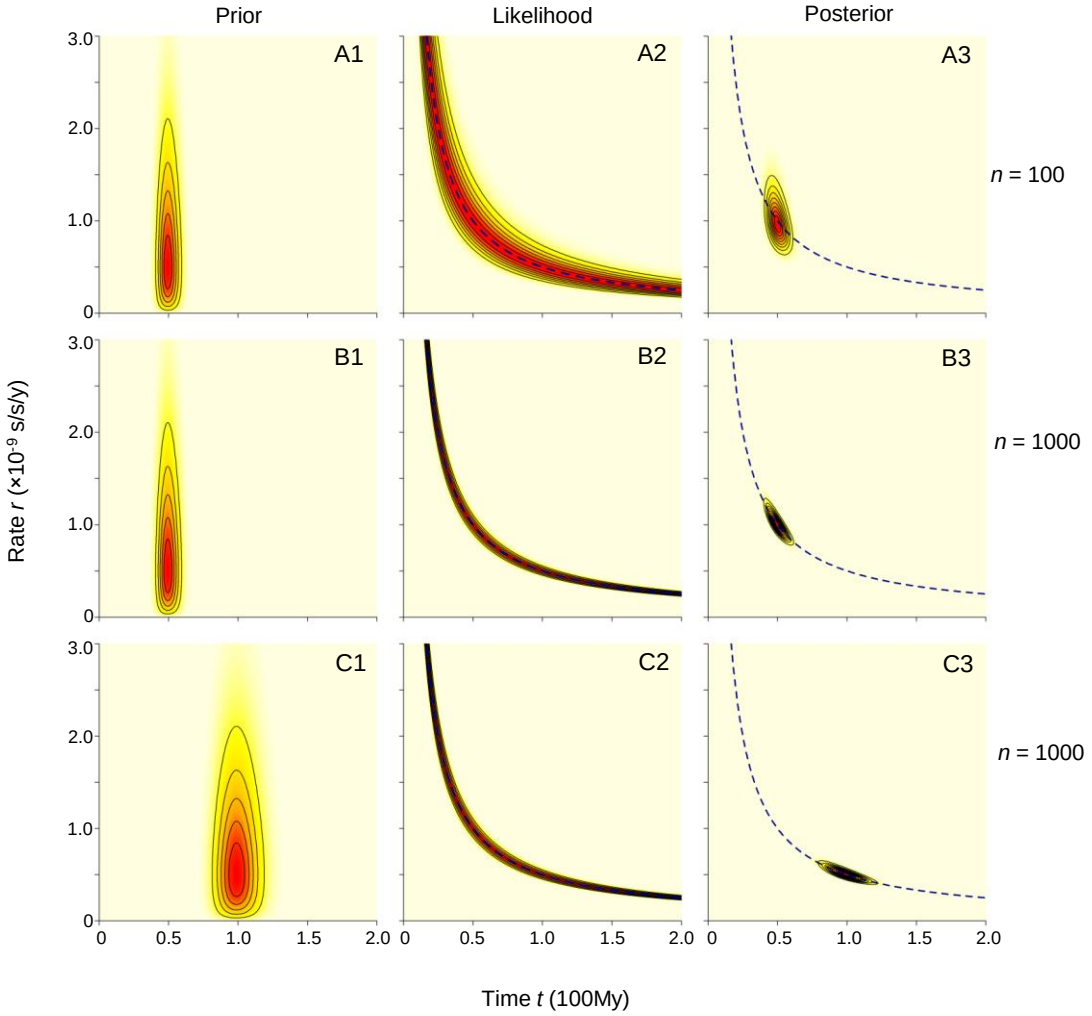

Figure 2.1: The prior, likelihood and posterior densities of time and rate for two data sets of a pairwise sequence alignment. In both data sets the true divergence time of the sequences is t = 50 Ma and the true rate is $r = 10^{-8}$ s/s/y meaning a true evolutionary distance d =1. In (A) the alignment consists of n = 100 sites with x = 55 differences while in (B) and (C) n = 1000 and x = 552. In (A) and (B) the joint prior is the product of two correct gamma priors $r \sim G(2, 2)$ and $t \sim G(100, 200)$, while in (C) the incorrect prior $t \sim G(100, 100)$ is used. In both data sets the likelihood is maximized along the $r = \frac{\hat{d}}{2t}$ line (blue dashed line). The Gaussian quadrature method (see Appendix A) and R scripts were used for the calculations.

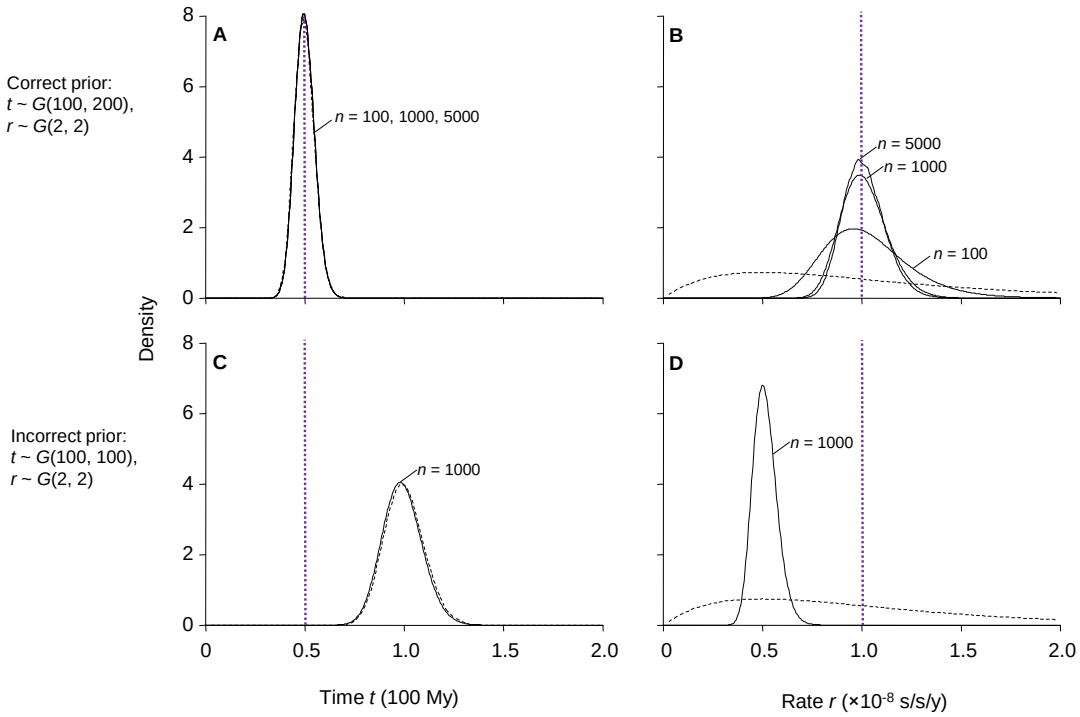

Figure 2.2: Marginal prior (dashed lines) and posterior (solid lines) distributions of time and rate for the data sets of Figure 2.1. A third data set with n = 5,000 sites is also considered here for the case of a correct prior. The marginal posterior of the rate for 5,000 sites is very similar to that for 1,000 sites indicating that inclusion of more molecular data is unlikely to reduce further the uncertainty in posterior estimates. The marginal posterior of the time is virtually the same for any alignment length and very similar to the prior because of the very informative time prior. Time and rate estimates are correct when a correct time prior is used (first line) but they are biased under an incorrect time prior (second line), indicating a strong prior influence. The vertical dotted lines indicate the true parameter values.

which cannot be reduced by adding more molecular data. Many fossil calibrations of good quality are in general helpful. Fossil calibrations are crucial since they exert a significant influence in the posterior estimates due to the confounding nature of time and rate. Objective representation of fossil information is challenging and extreme care should always be taken in the specification of calibration densities and in their impact in the posterior inference (Warnock, et al. 2015).

3 Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons

In the previous chapter we described some existing implementations of the Bayesian inference to study two important biological questions: times of species divergences and mode and strength of natural selection operated in a protein level. In this chapter we will present a new Bayesian method to estimate the nonsynonymous/synonymous rate ratio and evolutionary distance for a pair of protein-coding sequences and address problems of previous counting and maximum likelihood methods.

3.1 Counting and maximum likelihood methods

Several intuitive methods have been proposed to estimate the nonsynonymous/ synonymous rate ratio for pairwise sequence comparisons. Those methods, referred as counting or approximate methods, make ad-hoc treatments and are not based on a rigorous theory. Although they might differ in detail they all consist of three major steps: (i) count the numbers of synonymous and nonsynonymous sites in the alignment (ii) count the numbers of synonymous and nonsynonymous differences between the sequences and (iii) calculate the proportions of differences at the synonymous and nonsynonymous sites and correct for multiple substitutions using a standard evolutionary model.

Perler et al. (1980) and Miyata and Yasunaga (1980) developed the first counting methods to estimate the synonymous and nonsynonymous rates from a pair of sequences. Nei and Gojobori (1986) proposed a more efficient algorithm which produced similar estimates but it was simpler and thus soon became very popular. The method of Nei and Gojobori (NG) calculates the number of synonymous (S) and nonsynonymous (N) sites in the alignment by summing the respective counts over all codons and averaging between the two sequences. It then calculates the numbers of synonymous (S_d) and non synonymous (N_d) differences between the sequences by summing the respective counts across all codons. When a pair of codons has more than one nucleotide difference several evolutionary pathways exist and all of them are equally weighted. Then the proportions of synonymous (p_S) and nonsynonymous (p_N) differences are calculated by $p_S = S_d/S$ and $p_N = N_d/N$. Finally, the JC69 model is used to correct for multiple substitutions at a site with the synonymous

and nonsynonymous rates to be given by $d_S = -\frac{3}{4}\log\left(1-\frac{4}{3}p_S\right)$ and $d_N = -\frac{3}{4}\log\left(1-\frac{4}{3}p_N\right)$, respectively. The ω ratio and the evolutionary distance t are calculated by $\omega = d_N/d_S$ and $t = \frac{3S}{S+N}d_S + \frac{3N}{S+N}d_N$.

The method assumes equal codon frequencies, ignores the transition/transversion rate bias and does not correct properly for multiple substitutions at a site. The JC69 model assumes that a nucleotide at a site can change into any other nucleotide but in this method a nucleotide is allowed to only synonymous or nonsynonymous changes. This ad hoc treatment to correct for multiple substitutions does not introduce too much bias in small sequence divergences and failure to account for codon frequency bias and transition/transversion rate bias might be more important. However, in high sequence divergences the ad hoc correction can introduce significant positive bias, producing high ω values for large t (Yang and Nielsen 2000). Due to the structure of the genetic code a transition in the third codon position is more likely to be synonymous than a transversion is. Thus, ignoring the transition/transversion bias (the method actually assumes $\kappa = 1$) causes underestimation of S and overestimation of N which results in underestimation of the ω ratio. Li et al. (1985) developed a method to accommodate differences in the transition and transversion rates by classifying each codon position into 2-fold, 4-fold and nondegenerate classes. The degeneracy of a codon position is determined by the number of synonymous changes. However, Li's method is more complicated than the NG and gives similar estimates. The method was later improved by Li (1993) and Pamilo and Bianchi (1993), but the method of Ina (1995) was the first to fully account for κ in all steps of the estimation. The assumption of equal codon frequencies affects the calculation of synonymous and nonsynonymous sites leading to biased ω estimates. The direction and magnitude of the bias depends on the observed frequencies with higher departure from equality producing more bias.

Ina's method constitutes an improvement but, as with many of the previous methods, uses equal weighting of pathways when counting the differences, introducing bias towards 1 in the estimation of the ω ratio. The bias is more serious in high sequence divergences where codons are more likely to differ in two of three positions and thus multiple pathways are plausible. Pathways with more synonymous changes are more likely when $\omega < 1$ and using equal weighting underestimates S_d and overestimates N_d. Yang and Nielsen (2000) developed an iterative algorithm incorporating features of codon models and likelihood estimation (Goldman and Yang 1994). The method accounts for transition/transversion bias, codon frequency bias and uses appropriate weights for the different pathways according to their relative probabilities of occurrence. The HKY85 model of nucleotide substitution is used to

correct for multiple hits (Hasegawa, et al. 1985). The algorithm performs better than the previous counting methods but worse than the ML method and tends to overestimate ω at small sequence divergences and overestimate it at large divergences.

The heuristic counting methods make ad hoc treatments to deal with specific features of molecular evolution which are not rigorously justified and they seem to suffer mainly on the way they classify the sites into synonymous and nonsynonymous categories. Models of codon evolution (Goldman and Yang 1994; Muse and Gaut 1994) account for those factors and the ML method is used to estimate the parameters of the models. Factors like transition/transversion rate ratio, nonsynonymous/synonymous rate ratio and frequency codon bias can be taken into account by the codon model by incorporating them as parameters inside a substitution rate matrix. For example, the rate matrix of equation (2.5) incorporates explicitly those factors and their estimates can be obtained by applying standard maximum likelihood theory (Yang and Nielsen 1998). Correction for multiple hits is also performed automatically by the codon model. Moreover, the likelihood estimation is simpler than the complicated calculations required by the counting methods and more realistic assumptions can be modelled in a straightforward way based on a rigorous statistical theory. The ML method for pairwise sequence comparisons returns more reliable estimates of ω than the counting methods and only for very short sequences and high ω values some counting methods may be advantageous (Yang and Nielsen 2000). Furthermore, a great advance of the likelihood approach is that it can be used to estimate the ω ratio from multiple sequences accounting for their phylogenetic relationship (Goldman and Yang 1994).

MLEs of ω for thousand of genes are routinely calculated as descriptive statistics in genome-scale comparisons (Nielsen, et al. 2005; Ge, et al. 2008; Walters and Harrison 2010; Buschiazzo, et al. 2012; Gladieux, et al. 2013; Wang and Chen 2013). Although the ML method for pairwise comparisons is quick and produces reasonable estimates of ω and t for most data sets, it suffers from a few problems when the data sets are extreme. For example, the MLE of ω ($\hat{\omega}$) is 0 when the two compared sequences have only synonymous differences and ∞ when they have only nonsynonymous differences. Similarly, when the sequences are identical, the MLE $\hat{t}$ is 0 and $\hat{\omega}$ is not unique. When the sequences are very divergent $\hat{t}$ may be ∞.

Because of these infinite or undefined estimates, neither $\hat{\omega}$ nor $\hat{t}$ have finite means or variances. Extreme values of $\hat{\omega}$ and $\hat{t}$ are commonly encountered in genome-level comparisons of thousands of genes, and those estimates cause difficulties with the calculation of summary statistics (such as mean $\hat{\omega}$ and $\hat{t}$ across all genes in the genome). Furthermore, statistical theory establishes that the MLEs are asymptotically unbiased, meaning that in large samples the expectation of an estimate equals the true value of the

parameter. However, in small samples MLEs may suffer from substantial biases. For short alignments the ML method for pairwise sequence comparisons has been observed to overestimate ω when the sequence divergence is small with the bias decreasing for higher sequence divergences or longer alignments (dos Reis and Yang 2013d).

A statistical method which always produces finite and reasonable estimates for ω and t is thus desirable. In the next section we develop a Bayesian method to calculate the posterior means of ω and t for pairwise sequence comparisons. The advance of the Bayesian approach is that using appropriate priors on ω and t the estimates are shrinked away from the extreme values of 0 and ∞ and have well-defined means and variances. Moreover, a combination of computer simulation and real data analysis reveals better frequentistic properties for the Bayesian estimates than the MLEs. The new Bayesian method is computationally efficient and thus appropriate for genomic comparisons of protein-coding genes.

3.2 The new Bayesian approach

Assume that we have an alignment of two protein-coding sequences from two species and we are interested in estimating the ω ratio. We model the evolution of codon sequences as a continuous-time Markov process using the codon model of Yang and Nielsen (1998). The model incorporates explicitly the transition/transversion rate ratio, the nonsynonymous/ synonymous rate ratio and the codon frequencies and accounts for the structure of the genetic code. The instantaneous substitution rate from codon i to codon j ($i \neq j$) is given by equation (2.5). The likelihood function, that is the probability of the pairwise sequence alignment x given ω, t, κ is

$$f(x \mid \omega, t, \kappa) = \prod_{h=1}^{L_c} \pi_i P_{ij}(t), \tag{3.1}$$

where i and j are the observed codons in the two sequences at site h, π_i is the equilibrium frequency of codon i and L_c is the length of the alignment in codons. $P_{ij}(t)$ is the probability that the codon i is replaced by codon j after time t, where t is measured by the expected number of nucleotide substitutions per codon (see §2.2.2).

The joint posterior distribution of ω and t is given by

$$f(t, \omega \mid x) = \frac{1}{C} f(x \mid t, \omega) f(t, \omega), \tag{3.2}$$

where $f(t,\omega)$ is the joint prior on t and ω and $C = \int_0^\infty \int_0^\infty f(x \mid t, \omega) f(t, \omega) \mathrm{d}t \mathrm{d}\omega$ is the normalizing constant. To avoid calculations of high dimensional integrals we replace the parameter κ in (3.2) with its MLE $\hat{\kappa}$. If the two sequences are identical so that $\hat{\kappa}$ is not

unique, we fix it at 2. The codon frequency parameters are estimated by the observed nucleotide or codon frequencies (Goldman and Yang 1994). The joint prior $f(t,\omega)$ is constructed as the product of two independent gamma distributions

$$f(t,\omega) = G(t\,|\,1.1, 1.1) \times G(\omega\,|\,1.1, 2.2), \tag{3.3}$$

where $G(\theta|\alpha, \beta)$ is the gamma density with mean α/β and variance α/β^2. Here, the prior means of t and ω are 1 and 0.5, respectively. We use a shape parameter $\alpha = 1.1$, meaning that the priors on t and ω are quite diffuse. Note that the joint prior has a mode away from (0, 0) and the prior density decays to 0 as either ω or t approaches infinity, thus penalizing extreme values. As point estimates of ω and t we use their posterior means

$$\tilde{\omega} = \mathrm{E}(\omega\,|\,x) = \frac{1}{C}\int_0^\infty\int_0^\infty \omega\, f(x\,|\,t,\omega,\hat{\kappa})\, f(t,\omega)\mathrm{d}t\mathrm{d}\omega, \tag{3.4}$$

$$\tilde{t} = \mathrm{E}(t\,|\,x) = \frac{1}{C}\int_0^\infty\int_0^\infty t\, f(x\,|\,t,\omega,\hat{\kappa})\, f(t,\omega)\mathrm{d}t\mathrm{d}\omega. \tag{3.5}$$

The posterior variances and covariance of ω and t can be similarly defined and can be calculated by

$$Var(\omega\,|\,x) = E(\omega^2\,|\,x) - \left[E(\omega\,|\,x)\right]^2, \tag{3.6}$$

$$Var(t\,|\,x) = E(t^2\,|\,x) - \left[E(t\,|\,x)\right]^2, \tag{3.7}$$

$$Cov(\omega, t\,|\,x) = E(\omega t\,|\,x) - E(\omega\,|\,x)E(t\,|\,x). \tag{3.8}$$

Thus, six double integrals need to be computed, one for the normalizing constant C, and five for the different expectations in equations (3.4) - (3.8).

Consider the calculation of the normalizing constant C. All other integrals are calculated similarly. We write $g(t, \omega) = f(x|t, \omega)f(t, \omega)$. For numerical stability we set $h(t, \omega) = \exp\{\log g(t, \omega) - l_{max}\}$, where l_{max} is a constant chosen for scaling such as the maximum of $\log g(t, \omega)$ or the maximum log-likelihood. The normalizing constant then becomes

$$C = \exp(l_{\max})\int_0^\infty\int_0^\infty h(t,\omega)\mathrm{d}t\mathrm{d}\omega. \tag{3.9}$$

The Gaussian quadrature method is used to calculate all integrals numerically (Yang 2014). Gaussian quadrature uses Legendre polynomials to approximate any continuous integrand function $f(x, y)$ and calculates the integral as

$$\int_{-1}^{1}\int_{-1}^{1} f(x,y)\mathrm{d}x\mathrm{d}y \approx \sum_{i,j=1}^{n} w_i w_j\, f(x_i, y_j), \tag{3.10}$$

where the weights w_i, w_j and the points x_i, x_j are predetermined given the total number of points n (Appendix A). Here, the limits of the integrals are 0 and ∞ and thus a transformation to map the $(0, \infty)$ limits to $(-1, 1)$ is necessary. If the integrand function $f(x, y)$ is highly

concentrated in a very small interval and a few points are used those will more likely miss the spike in the integrand and the approximation will be poor. The idea behind the transformation is to use a probability density function (PDF) that has a similar shape to the integrand $g(t, \omega)$ and use its cumulative distribution function (CDF) to transform the integrand. In that case the new transformed integrand will be nearly flat and a good approximation will be possible with just a few points. Note that if the chosen PDF matches exactly the $g(t, \omega)$, the new integrand after the transformation will be perfectly flat.

We use the logistic distribution to perform the mapping $(0, \infty) \rightarrow (-1, 1)$. For any random variable $x \sim \text{Logistic}(\mu, \sigma)$ the CDF is $F_L(x) = \frac{1}{1+\mathrm{e}^{-(x-\mu)/\sigma}}$. Let $x_1 = \log t \sim \text{Logistic}(\mu_1, \sigma_1)$ and $x_2 = \log\omega \sim \text{Logistic}(\mu_2, \sigma_2)$. Thus, for equation (3.9), we use the following change of variables:

$$z_1 = 2F_L(x_1) - 1 \Rightarrow t = \exp\left\{\mu_1 + \sigma_1 \log\frac{1+z_1}{1-z_1}\right\}, \tag{3.11}$$

$$z_2 = 2F_L(x_2) - 1 \Rightarrow \omega = \exp\left\{\mu_2 + \sigma_2 \log\frac{1+z_2}{1-z_2}\right\}. \tag{3.12}$$

Thus, the integral C becomes

$$C = \exp(l_{\max}) \int_{-1}^{1}\int_{-1}^{1} r(z_1, z_2)\mathrm{d}z_1\mathrm{d}z_2 \approx \exp(l_{\max}) \sum_{i,j=1}^{n} w_i w_j r(z_{1_i}, z_{2_j}), \tag{3.13}$$

where

$$r(z_1, z_2) = h(t, \omega)\frac{2t\sigma_1}{1-z_1^2}\frac{2\omega\sigma_2}{1-z_2^2}, \tag{3.14}$$

where t and ω are given by (3.11) and (3.12), respectively. Figure 3.1 shows the integrand function $r(z_1, z_2)$ after the logistic transformation for two sequences with only nonsynonymous differences. The same transformation is applied to all integrals in equations (3.4)-(3.8). Thus, we have

$$\begin{aligned}
\mathrm{E}(\omega \mid x) &\approx \frac{1}{A}\sum_{i,j=1}^{n} w_i w_j \omega_i r(z_{1_i}, z_{2_j}),\\
\mathrm{E}(t \mid x) &\approx \frac{1}{A}\sum_{i,j=1}^{n} w_i w_j t_j r(z_{1_i}, z_{2_j}),\\
\mathrm{E}(\omega^2 \mid x) &\approx \frac{1}{A}\sum_{i,j=1}^{n} w_i w_j \omega_i^2 r(z_{1_i}, z_{2_j}),\\
\mathrm{E}(t^2 \mid x) &\approx \frac{1}{A}\sum_{i,j=1}^{n} w_i w_j t_j^2 r(z_{1_i}, z_{2_j}),\\
\mathrm{E}(t\omega \mid x) &\approx \frac{1}{A}\sum_{i,j=1}^{n} w_i w_j \omega_i t_j r(z_{1_i}, z_{2_j}),
\end{aligned} \tag{3.15}$$

where $A = C\exp(-l_{max})$. The constant term $\exp(l_{max})$ cancels and is not involved in the calculations of the expectations.

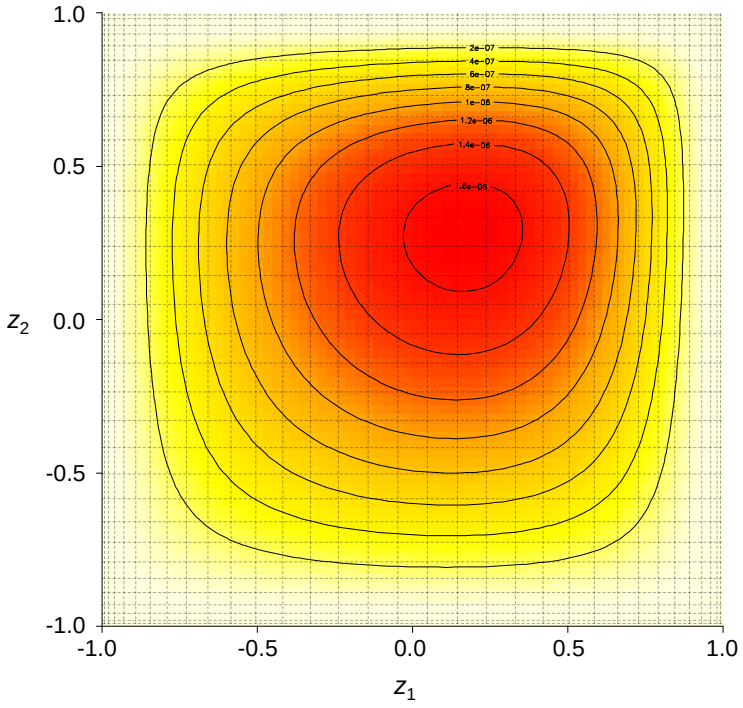

Figure 3.1: Integrand function of t and ω after logistic transformation for the calculation of the normalizing constant (equation 3.9). The data is an alignment of two sequences of 100 codons with only nonsynonymous differences (S = 73.2, N = 226.8, S_d = 0, N_d = 40). The values of t and ω are given from (3.11) and (3.12), respectively, while the transformed integrand function is according to (3.14). The new integrand is not spiky and thus the integral can be calculated reliably using a small number of points in the Gaussian quadrature method. The grid shows the points at which the new integrand function $r(z_1, z_2)$ is evaluated. For each dimension n = 32 were used implying 32 × 32 = 1024 evaluations of the integrand.

The Bayesian calculation is performed after the MLEs are obtained. Thus, if both $\hat{\omega}$ and $\hat{t}$ are finite, away from 0 and the observed p_S and p_N are < 0.74, we set $\mu_1 = \log \hat{t}$, $\mu_2 = \log \hat{\omega}$, $\sigma_1 = \left(\frac{1}{\hat{t}}\right)\sqrt{\hat{V}(\hat{t})}$ and $\sigma_2 = \left(\frac{1}{\hat{\omega}}\right)\sqrt{\hat{V}(\hat{\omega})}$. The variances $\hat{V}(\hat{t})$ and $\hat{V}(\hat{\omega})$ are estimated using the Nei and Gojobori (1986) method (Appendix B). Because this method uses the JC69 model to correct for multiple hits, the use of 0.74 as an upper limit for the p_S and p_N guarantees an adequate estimation of $\hat{V}(\hat{t})$ and $\hat{V}(\hat{\omega})$.

In all other cases, we find numerically the point ($\bar{t}$, $\bar{\omega}$) that maximizes $\log\{g(t, \omega)\}$. We calculate the Hessian matrix at this point using the second-order difference method and

we use the inverse of the Hessian to estimate the variances $V(\bar{t})$ and $V(\bar{\omega})$. Then, we set $\mu_1 = \log \bar{t}$, $\mu_2 = \log \bar{\omega}$, $\sigma_1 = \left(\frac{1}{\bar{t}}\right)\sqrt{\hat{V}(\bar{t})}$ and $\sigma_2 = \left(\frac{1}{\bar{\omega}}\right)\sqrt{\hat{V}(\bar{\omega})}$. Note that because the joint prior has a mode, log(g) always has a mode and thus ($\bar{t}$, $\bar{\omega}$) is always away from (0, 0). Although this approach can be used in all cases, no matter the values of p_S, p_N, $\hat{\omega}$ and $\hat{t}$, optimization of log(g) is computationally expensive.

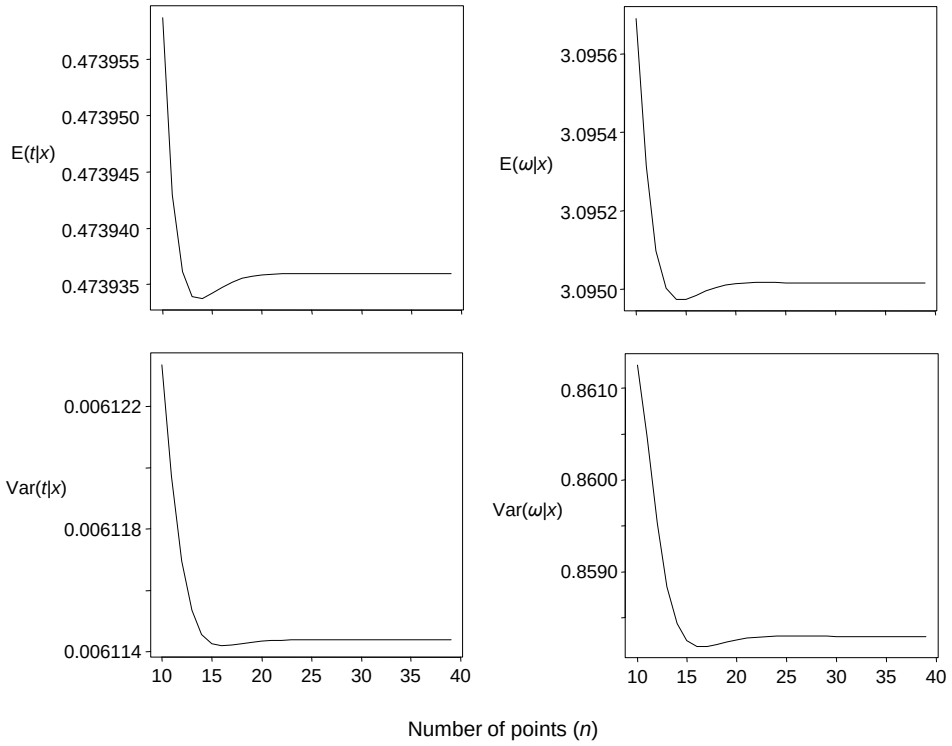

Figure 3.2: Estimated posterior mean and variance of t and ω according to the number of points used in the Gaussian quadrature method. The data are the same as in Figure 3.1 and the joint prior on t and ω is given by equation (3.3). The estimates are stable for $n > 25$.

Except for the posterior means E(t|x), E(ω|x), variances Var(t|x) and Var(ω|x) and covariance Cov(t,ω|x), we also calculate the posterior probability for $\omega > 1$ given by

$$P(\omega > 1 | x) = \frac{1}{C}\int_0^\infty\int_1^\infty f(x | t, \omega, \hat{\kappa}) f(t, \omega) \mathrm{d}t\mathrm{d}\omega . \qquad (3.16)$$

This serves as a Bayesian alternative to the LRT of the null hypothesis H_0: $\omega = 1$ with alternative H_1: $\omega > 1$ to test for positive selection (indicated by $\omega > 1$). To calculate the $P(\omega>1|x)$ we used similar techniques to those described above (Appendix C).

The same number of points n for both parameters t and ω was used in the Gaussian quadrature method for simplicity. With $n = 32$, each sum in equation (3.15) requires 32×32

= 1024 evaluations of the $r(z_1, z_2)$ function. The use of more points increases the computational time radically since evaluation of $r(z_1, z_2)$ requires calculation of the likelihood which is computationally expensive. Tests suggested that using 32 points achieves high accuracy. Figure 3.2 shows the estimates of E(t|x), E(ω|x), Var(t|x) and Var(ω|x) for an alignment of two sequences with only nonsynonymous differences using different numbers of points in the Gaussian quadrature. The estimates become stable when more than 25 points are used.

The Bayesian calculation of ω and t was implemented in the CODEML program (Yang 2007).

3.3 Simulations

3.3.1 Performance of the Bayesian method in five different data sets

To highlight differences among the ML and Bayesian methods we consider five different scenarios in which the numerical calculations of the integrals may differ. For each case we simulated an alignment of two sequences of 100 codons in length, with different numbers of synonymous and nonsynonymous differences. The log-likelihood and log-posterior surfaces for the five cases are shown in Figure 3.3.

Case 1: ($S_d > 0$, $N_d > 0$). This is the most common case with both synonymous and nonsynonymous differences observed. The data are quite informative about ω and t and the posterior distribution resembles the likelihood (Figure 3.3 A' and A). In this data set, we have S = 73.7, N = 226.3, S_d = 18.5, N_d = 6.5. The MLEs are $\hat{t}$ = 0.30 and $\hat{\omega}$ = 0.11 while the posterior means are $\tilde{t}$ = 0.31 and $\tilde{\omega}$ = 0.13, very close to the MLEs.

Case 2: ($S_d = N_d = 0$). In this case, the two sequences are identical with S = 73.3, N = 226.7 and $S_d = N_d = 0$. The likelihood is maximized along the $t = 0$ line and when $t = 0$, ω has no effect on the likelihood; therefore ω has no unique MLE (Figure 3.3 B). However, the posterior has a single mode and the posterior means are $\tilde{t}$ = 0.011 and $\tilde{\omega}$ = 0.496 (Figure 3.3 B'). Note that since the data are uninformative about ω the posterior mean is almost equal to the prior mean. Moreover, the posterior mean is far from the posterior mode, because the joint posterior is highly skewed. Note also that the posterior means refer to the marginal posteriors and thus the point ($\tilde{t}$, $\tilde{\omega}$) may differ from the mean of the joint posterior (depending on the correlation of t and ω).

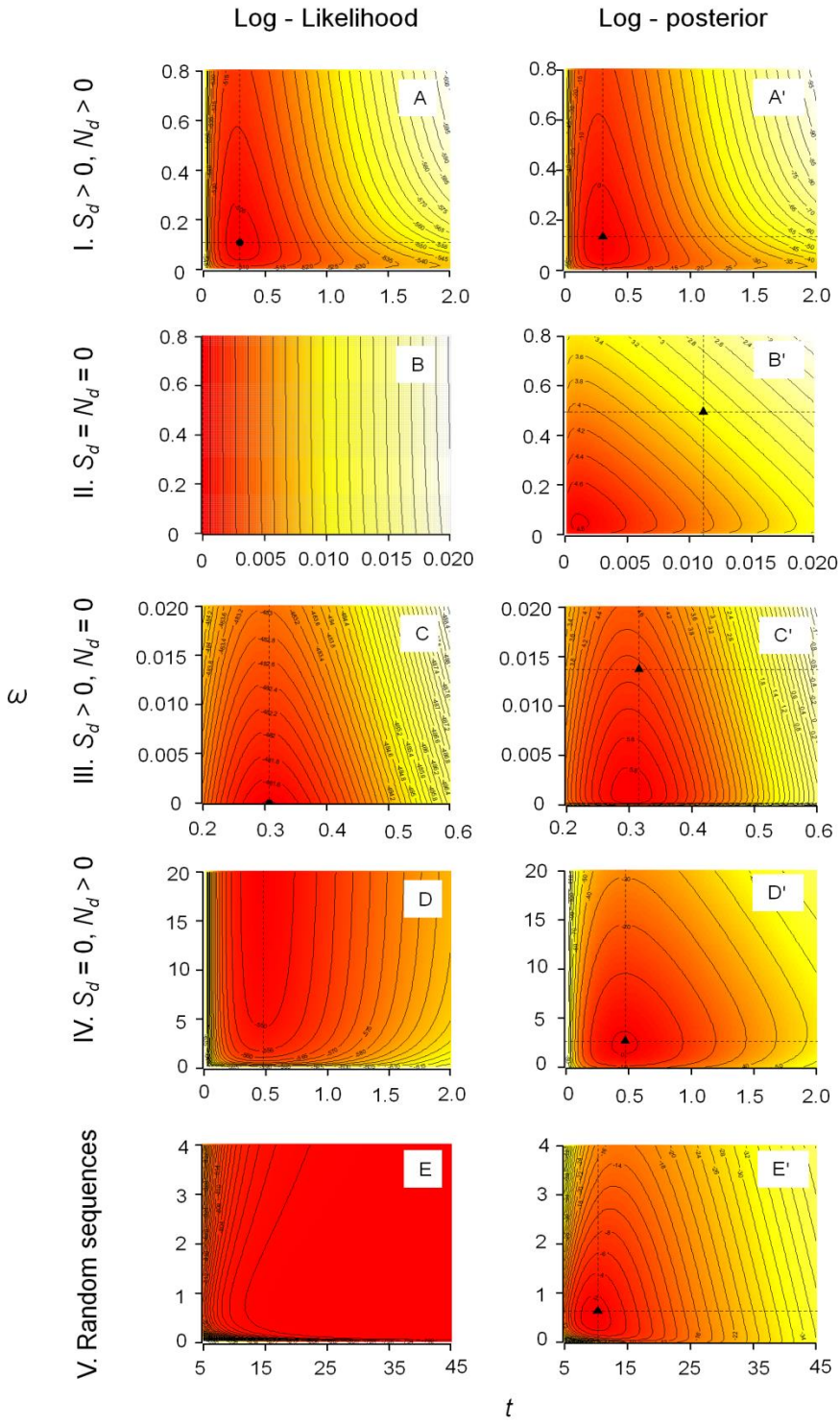

Figure 3.3: Contour plots of log-likelihood (A-E) and log-posterior (A'-E') distributions of ω and t for five artificial alignments of two sequences with 100 codons. The black disc in A-E indicates the MLE while the black triangle in A'-E' indicates the mode of the joint posterior. The five cases are: (i) normal sequences with both synonymous and nonsynonymous differences (A, A'), (ii) identical sequences (B, B'), (iii) sequences with only synonymous differences (C, C'), (iv) sequences with only nonsynonymous differences (D, D') and (v) highly divergent sequences (E, E').

Case 3: ($S_d > 0$, $N_d = 0$). The two sequences differ by only synonymous changes. In this data set $S = 74.4$, $N = 225.6$, $S_d = 24$ and $N_d = 0$. The MLEs are $\hat{t} = 0.306$ and $\hat{\omega} = 0$ (Figure 3.3 C). However, the posterior has a mode away from $\omega = 0$ and the posterior means are $\tilde{t} = 0.316$ and $\tilde{\omega} = 0.014$ (Figure 3.3 C').

Case 4: ($S_d = 0$, $N_d > 0$). The two sequences differ by only nonsynonymous changes. In this data set $S = 73.2$, $N = 226.8$, $S_d = 0$, $N_d = 40$. We have $\hat{t} = 0.48$ and the likelihood surface increases asymptotically along the $t = 0.48$ line, so that $\hat{\omega} = \infty$ (Figure 3.3 D). However, the posterior has a well-defined mode and thus $\tilde{t} = 0.47$ and $\tilde{\omega} = 3.1$ (Figure 3.3 D').

Case 5: ($S_d \gg 0$, $N_d \gg 0$). The two sequences are so highly divergent that they practically look like random sequences ($S = 75.9$, $N = 224.1$, $S_d = 75$, $N_d = 175$). Here, the likelihood increases with the increase of both t and ω, with the MLEs at $\hat{t} = \infty$ and $\hat{\omega} = \infty$ (Figure 3.3 E). In the Bayesian analysis, the prior penalizes extreme values and thus the posterior means are $\tilde{t} = 10.31$ and $\tilde{\omega} = 0.72$ (Figure 3.3 E'). Note that the posterior mean of ω is close to the prior mean. Because the sequences are very divergent there is too much noise and the sequences are practically uninformative about ω. For example, it is hard to tell whether a nonsynonymous difference is due to a high ω ratio or because the divergence between the two sequences is high.

These five cases demonstrate how the prior influences the posterior depending on whether the data are informative or not. The posterior means of t and ω are finite for all five cases, whereas the MLEs are not. The mean square error (MSE) of an estimator $\hat{\theta}$ for an unknown parameter θ is defined as $\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + [\mathrm{E}(\hat{\theta}) - \theta]^2$ and is used to assess the quality of an estimator in terms of its variation and degree of bias. Because the MLEs of t and ω may be infinite, their MSEs are ∞ as well. In contrast, the MSEs of the posterior means are always defined. Thus, in this sense, the Bayesian estimates of t and ω have better Frequentistic properties compared to their MLEs counterparts.

In the following section we perform a thorough simulation analysis to study the statistical properties of the new Bayesian estimators of ω and t and compare them with the traditional MLEs.

3.3.2 Analysis of simulated data

We used the program EVOLVER from the PAML package (Yang 2007) to simulate pairwise sequence alignments of length $L_c = 500$ codons. We used $t = 0.1, 0.5, 1, 5$ and $\omega = 0.01, 0.1, 0.5, 2$, (16 combinations) with transition/ transversion rate ratio $\kappa = 2$ and equal codon frequencies (1/61) to simulate the data sets. The number of replicates was 10,000.

Then we used the CODEML program (Yang 2007) to analyze the simulated data sets using both ML and the new Bayesian method assuming equal codon frequencies (Fequal model). In the Bayesian method the same prior (equation 3.3) was used for all data sets.
Figures 3.4 and 3.5 show the histograms (smoothed densities) of posterior mean estimates and MLEs of ω and t. As we see in Figure 3.4 the ML and Bayesian estimates of ω are virtually identical for all combinations of ω = 0.1, 0.5 and t = 0.5, 1. However, for ω = 0.01, Bayesian estimates of ω are shifted to the right (too large) for all t values. This is because the prior for ω has a mean of 0.5 and affects the posterior estimates. For ω = 2, posterior means of ω are shifted to the left (too small) because of the prior. In general, both methods behave best (histograms are more concentrated around the true values) for intermediate distances (t = 0.5, 1), because sequences of moderate divergences are the most informative. Similar patterns are observed for the distance estimates (Figure 3.4). For t = 0.5, 1 the Bayesian estimates are almost identical to the MLEs, but for t = 0.1 they are slightly shifted to the right (too large) and for t = 5 they are shifted to the left (too small).

Table 3.1 and 3.2 contain the means of the Bayesian and ML estimates, the square root of the MSE ($\sqrt{MSE}$), and the 2.5% and 97.5% percentiles of estimates from the 10,000 replicate data sets. The descriptive statistics for the ML method have been calculated after removing the infinite estimates. For very similar (t = 0.1) and very divergent (t = 5) sequences, the prior has a noticeable impact. For example, when t = 0.1 the mean of Bayesian estimates of ω is 0.02 when the true ω = 0.01 and is 1.591 when the true ω = 2.0. The means of the MLEs (0.011 and 2.365, respectively) are in comparison closer to the true values than the means of the Bayesian estimates. However, the means of the MLEs are calculated after data sets in which $\hat{\omega} = \infty$ are excluded, but the same data sets are included in the calculation of the Bayesian estimates. Estimates of distance show similar patterns (Table 3.2). Furthermore, when t and ω are small or intermediate, ML and Bayesian methods have similar MSE, but for large ω and t the Bayesian estimates have smaller MSE indicating that in those cases Bayesian estimates are preferable to the MLEs.

We also tested for positive selection, indicated by $\omega > 1$. In the maximum likelihood analysis a LRT is used to compare H_0: $\omega = 1$ against H_1: $\omega > 1$, at the 5% significance level. In the Bayesian analysis positive selection is inferred when $P(\omega > 1 \mid x) > 0.95$. For the true ω = 0.01, 0.1, 0.5, no datasets were found to be under positive selection by either method. When the true ω = 2 and t = 0.5, 1, 5, both methods correctly detect positive selection in almost 100% of the replicate data sets, so that the power to detect positive selection is high in both methods but with the LRT to be slightly more powerful (Table 3.1). When ω = 2 and t = 0.1 the Bayesian and ML methods detect positive selection in 35% and 61% of data sets. In this case, given the short sequence distance, the prior has quite some impact on the ability

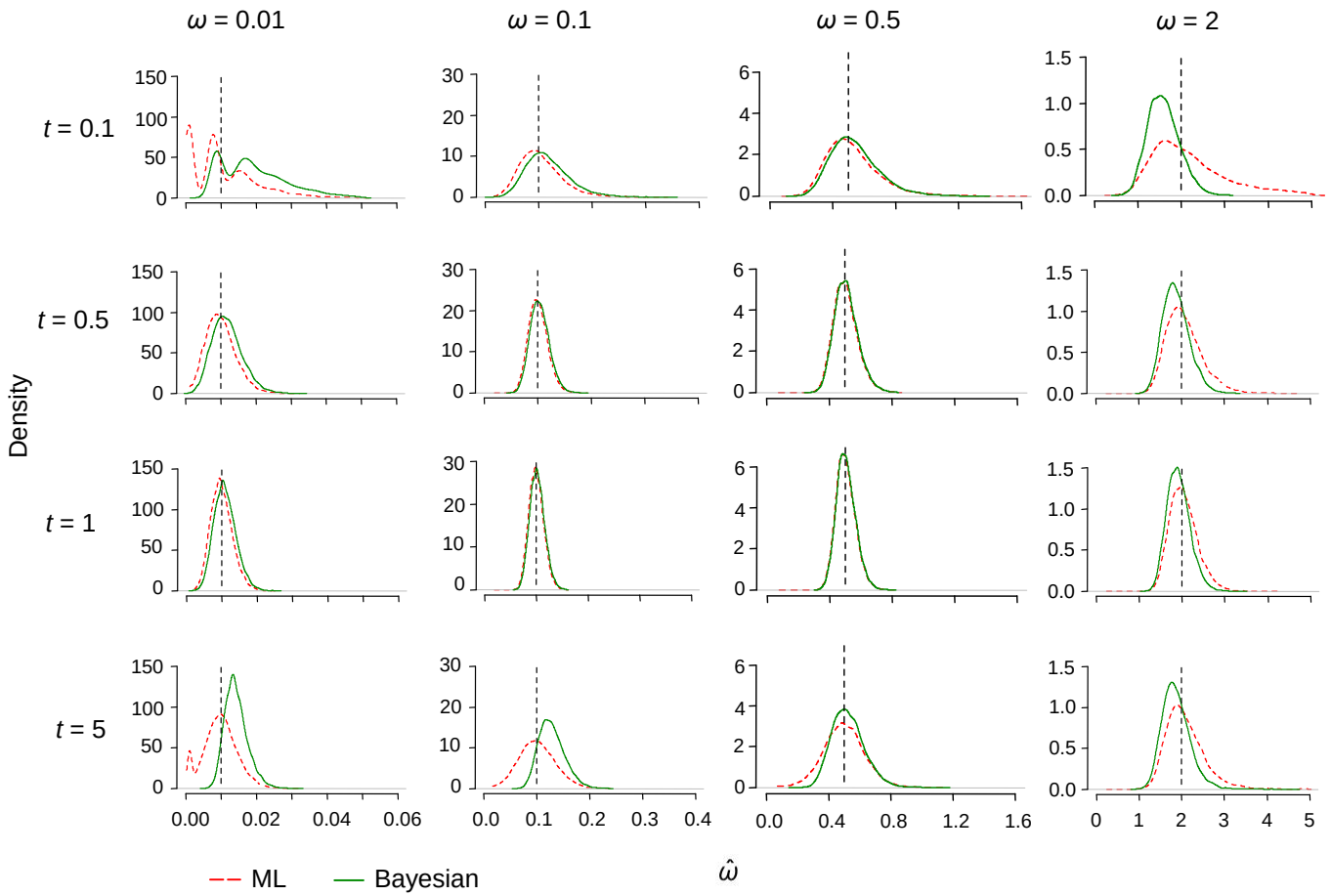

Figure 3.4: Kernel densities (smoothed histograms) of MLEs (dashed red) and Bayesian posterior means (solid green) for ω in simulated data sets. The true values of ω and t are shown on top and left of the plots, respectively. The sequence length is 500 codons. The number of replicates is 10,000. The vertical dashed lines correspond to the true values of ω. Independent gamma priors are used $\omega \sim G(1.1, 2.2)$ and $t \sim G(1.1, 1.1)$.

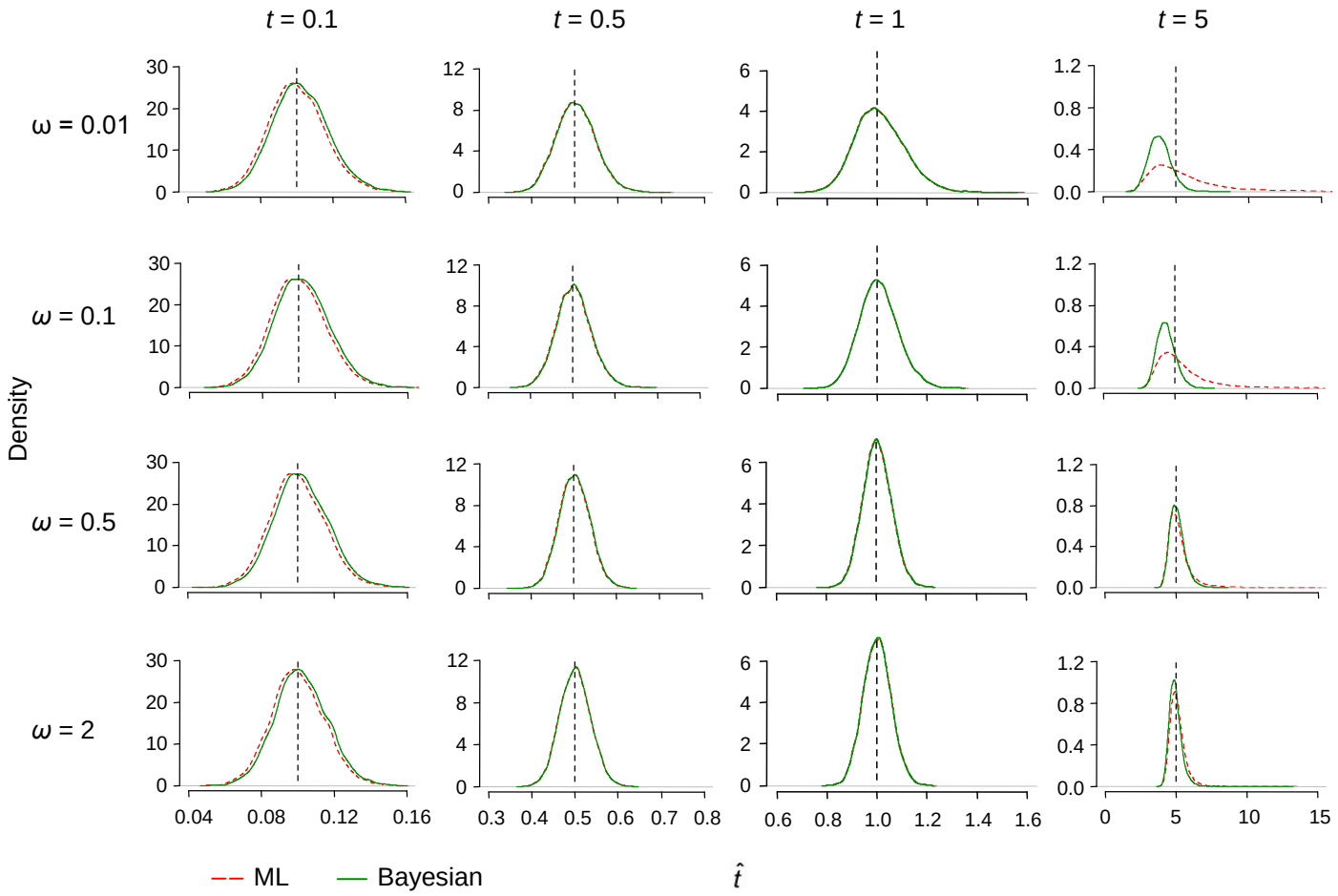

Figure 3.5: Kernel densities (smoothed histograms) of MLEs (dashed red) and Bayesian posterior means (solid green) for t in simulated data sets. See legend of Figure 3.4 for details.

Table 3.1: Summary statistics of Bayesian (top, bold) and ML (bottom) estimates of ω from 10,000 simulated data sets

	ω = 0.01					ω = 0.1				ω = 0.5					ω = 2					
	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	N_0	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	N_∞	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	N_∞	P_+
t = 0.1	**0.020**	**0.014**	**0.007**	**0.044**	**0**	**0.118**	**0.045**	**0.052**	**0.214**	**0.543**	**0.160**	**0.301**	**0.904**	**0**	**1.591**	**0.546**	**0.966**	**2.359**	**0**	**35.1**
	0.011	0.009	0	0.033	2861	0.103	0.039	0.041	0.194	0.528	0.172	0.278	0.936	0	2.365	1.484	1.015	5.626	3	60.7
t = 0.5	**0.012**	**0.005**	**0.005**	**0.021**	**0**	**0.104**	**0.018**	**0.072**	**0.141**	**0.511**	**0.076**	**0.379**	**0.677**	**0**	**1.878**	**0.329**	**1.360**	**2.543**	**0**	**98.3**
	0.010	0.004	0.003	0.019	15	0.101	0.018	0.069	0.138	0.506	0.076	0.374	0.674	0	2.064	0.424	1.409	3.031	0	98.9
t = 1	**0.011**	**0.003**	**0.006**	**0.018**	**0**	**0.102**	**0.014**	**0.076**	**0.132**	**0.506**	**0.062**	**0.397**	**0.637**	**0**	**1.922**	**0.278**	**1.466**	**2.497**	**0**	**99.9**
	0.010	0.003	0.005	0.017	0	0.100	0.014	0.075	0.130	0.503	0.062	0.393	0.635	0	2.038	0.326	1.508	2.764	0	100
t = 5	**0.014**	**0.005**	**0.009**	**0.022**	**0**	**0.129**	**0.038**	**0.089**	**0.183**	**0.526**	**0.109**	**0.348**	**0.755**	**0**	**1.876**	**0.374**	**1.331**	**2.642**	**0**	**97.4**
	0.010	0.005	0	0.019	370	0.101	0.034	0.037	0.171	0.515	0.981	0.226	0.762	44	2.120	1.398	1.400	3.228	0	98.6

Note.− Data were analyzed assuming equal codon frequencies (Fequal model). Results for ML have been calculated after removing infinite estimates. For ω = 0.1, there were no data sets with 0 or infinite estimates. N_0 is the number of replicates with $\hat{\omega} = 0$, whereas N_∞ is the number of replicates with $\hat{\omega} = \infty$. P_+ is the proportion of replicates with significant evidence for positive selection indicated by $P(\omega > 1 \mid x) > 0.95$ in the Bayesian method or by a significant LRT at the 5% level (one-sided with critical value 2.71) in the likelihood method.

Table 3.2: Summary statistics of Bayesian (top, bold) and ML (bottom) estimates of t from 10,000 simulated data sets

	t = 0.1				t = 0.5				t = 1				t = 5				
	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	N_∞
ω = 0.01	**0.102**	**0.015**	**0.074**	**0.134**	**0.504**	**0.045**	**0.421**	**0.596**	**1.013**	**0.100**	**0.837**	**1.223**	**3.910**	**1.322**	**2.600**	**5.506**	**0**
	0.100	0.015	0.072	0.132	0.503	0.045	0.419	0.595	1.011	0.100	0.836	1.222	7.572	8.922	2.676	43.744	244
ω = 0.1	**0.102**	**0.015**	**0.075**	**0.133**	**0.503**	**0.041**	**0.427**	**0.587**	**1.007**	**0.077**	**0.865**	**1.171**	**4.406**	**0.869**	**3.317**	**5.795**	**0**
	0.100	0.015	0.073	0.131	0.502	0.041	0.425	0.585	1.006	0.077	0.864	1.170	5.629	2.700	3.373	11.506	24
ω = 0.5	**0.102**	**0.015**	**0.075**	**0.132**	**0.503**	**0.036**	**0.436**	**0.574**	**1.004**	**0.057**	**0.895**	**1.118**	**5.158**	**1.469**	**4.249**	**6.368**	**0**
	0.100	0.015	0.073	0.130	0.501	0.036	0.434	0.572	1.002	0.057	0.894	1.116	5.440	2.601	4.228	7.979	43
ω = 2	**0.102**	**0.015**	**0.075**	**0.131**	**0.501**	**0.035**	**0.434**	**0.571**	**1.001**	**0.056**	**0.895**	**1.112**	**4.988**	**0.737**	**4.274**	**6.035**	**0**
	0.100	0.014	0.073	0.129	0.500	0.035	0.433	0.571	1.002	0.056	0.895	1.114	5.119	0.726	4.323	6.401	3

Note.− Data were analyzed assuming equal codon frequencies (Fequal model). Results for ML have been calculated after removing the infinite estimates. For t = 0.1, 0.5 and 1, there were no data sets with 0 or infinite estimate. N_∞ is the number of replicates with $\hat{\omega} = \infty$.

of the Bayesian method to detect selection. In particular, the prior mean (ω = 0.5) is smaller than the true value (ω = 2), and thus $\tilde{\omega}$ is shrunk away from 1.

3.4 Analysis of mammalian and bacterial data

We applied both ML and Bayesian methods to estimate ω and t for pairwise alignments of protein-coding genes from four mammalian species (human, chimpanzee, mouse, and rat) and from three bacterial strains (*Escherichia coli O157:H7*, *E. coli K-12* and *Salmonella typhimurium LT2*). The mammalian data set is a subset of the data analyzed by dos Reis et al. (2012). The data set consists of: 14,218 genes from the human and chimpanzee, with the sequence length ranging from 39 to 8,797 codons; 14,631 genes from the human and mouse with the sequence length from 13 to 8,787 codons; and 13,371 genes from the mouse and rat with the sequence length from 14 to 7,798 codons. The protein-coding sequences from the genomes of *E. coli O157:H7*, *E. coli K-12* and *S. typhimurium LT2* were downloaded from GenBank (accession numbers: U_00096, NC_002655 and NC_003197 respectively). Orthologous genes among the three genomes were identified by using the program BLAT (Kent 2002) to extract the best reciprocal hits. Only orthologs present in all three genomes are used. The bacterial data set consists of 2,631 genes from each strain, with the sequence length ranging from 20 to 1,485 codons. Codons involving alignment gaps and ambiguity nucleotides were removed prior to analyses. Moreover, genes with sequence length of 50 codons or less were excluded from the analysis. The number of genes analyzed in each comparison is reported in Table 3.3 and Figure 3.6. In all analyses, the codon frequencies were estimated by using the observed codon frequencies in the genes (the F61 model).

3.4.1 Analysis of the mammalian data set

We conducted three sets of pairwise comparisons: human versus chimpanzee, human versus mouse, and mouse versus rat. Figure 3.6 shows the distributions (smoothed histograms) of posterior means and the MLEs of t and ω in those comparisons. In the human–chimpanzee comparison, the Bayesian ω estimates are slightly shifted to the right compared with the MLEs for low ω values and shifted to the left for high ω values. The mean, median, 25% and 75% percentiles of the Bayesian estimates are 0.369, 0.320, and (0.180, 0.500) whereas those of the MLEs are 0.307, 0.193, and (0.062, 0.411) (Table 3.3). The human and chimpanzee genes are very similar and the patterns are similar to those observed in computer simulation for low t values. Moreover, there are 377 and 2507 gene

alignments in which $\hat{t} = 0$ and $\hat{\omega} = 0$, respectively, as well as 2 and 423 alignments where $\hat{t} = \infty$ and $\hat{\omega} = \infty$, respectively. The Bayesian method does not produce any such extreme estimates. The number of genes in which the ω estimate is greater than 1 is 1121 for ML and 299 for the Bayesian method (Table 3.4). The discrepancy is the result of two effects, a short evolutionary distance and a short sequence length, both indicating a lack of information and greater influence from the prior. Genes with $\hat{\omega} > 1$ tend to be small (median sequence length 313 codons, compared with 454 codons for all genes). For example, one gene among those 1121 with $\hat{\omega} > 1$ has $\hat{\omega} = 1.22$ (95% Confidence interval - CI 0.37 to 4.01) and posterior mean $\tilde{\omega} = 0.93$ (95% Credibility interval - CI 0.36 to 2.43). This gene has a length of 262 codons and has a small evolutionary distance with $\hat{t} = 0.043$ (95% CI 0.024 to 0.077) and $\tilde{t} = 0.047$ (95% CI 0.027 to 0.082), and thus the prior has an impact. Another gene has $\hat{\omega} = 1.27$ (95% CI 0.75 to 2.16) and $\tilde{\omega} = 1.13$ (95% CI 0.60 to 2.13). This gene is 257 codons in length and the ML and Bayesian distance estimates are 0.17 (95% CI 0.13 to 0.24) and 0.18 (95% CI 0.13 to 0.24) respectively. The second gene has a similar length to the first but because the sequence distance is greater, the prior is much less important. In a third gene, of length 1019 codons, the MLEs are $\hat{t} = 0.041$ (95% CI 0.030 to 0.056) and $\hat{\omega} = 1.27$ (95% CI 0.77 to 2.07), compared with the Bayesian estimates $\tilde{t} = 0.042$ (95% CI 0.031 to 0.057) and $\tilde{\omega} = 1.13$ (95% CI 0.59 to 2.14). In this case the effect of the prior is unimportant, because the gene is long.

Among the 1121 genes with $\hat{\omega} > 1$ only 78 have statistically significant evidence of positive selection based on the LRT ($\alpha = 5\%$) (Table 3.4). All the 78 genes have posterior mean $\tilde{\omega} > 1$. Moreover, out of them, three showed strong evidence of positive selection in the Bayesian analysis, with $P(\omega > 1 \mid x) > 0.95$ (Table 3.4). The difference (3 vs. 78 genes) in the number of genes found to be under positive selection between the ML and the Bayesian method is consistent with the general expectation that the likelihood ratio test tends to reject the null more readily than the Bayesian analysis. It is also consistent with the results observed in the computer simulations for $t = 0.1$ and $\omega = 2$. Note that the 3 genes significant in the Bayesian analysis have fairly large sequence divergences, with $\hat{t} \approx 0.1$, while the other 75 genes (for which the LRT is significant but the Bayesian evidence is not strong) have highly similar sequences, with $\hat{t} < 0.07$ (and median 0.021).

In the human-mouse comparison, the ML and Bayesian estimates are very similar. The sequence divergence is intermediate, the data are informative, and thus the prior does not have a noticeable impact. There are very few cases where the MLEs are extreme (0 or ∞) (Table 3.3). Also, the number of genes with ω estimates > 1 is nearly the same between the two methods (6 vs. 7) and the same two genes show significant evidence for positive

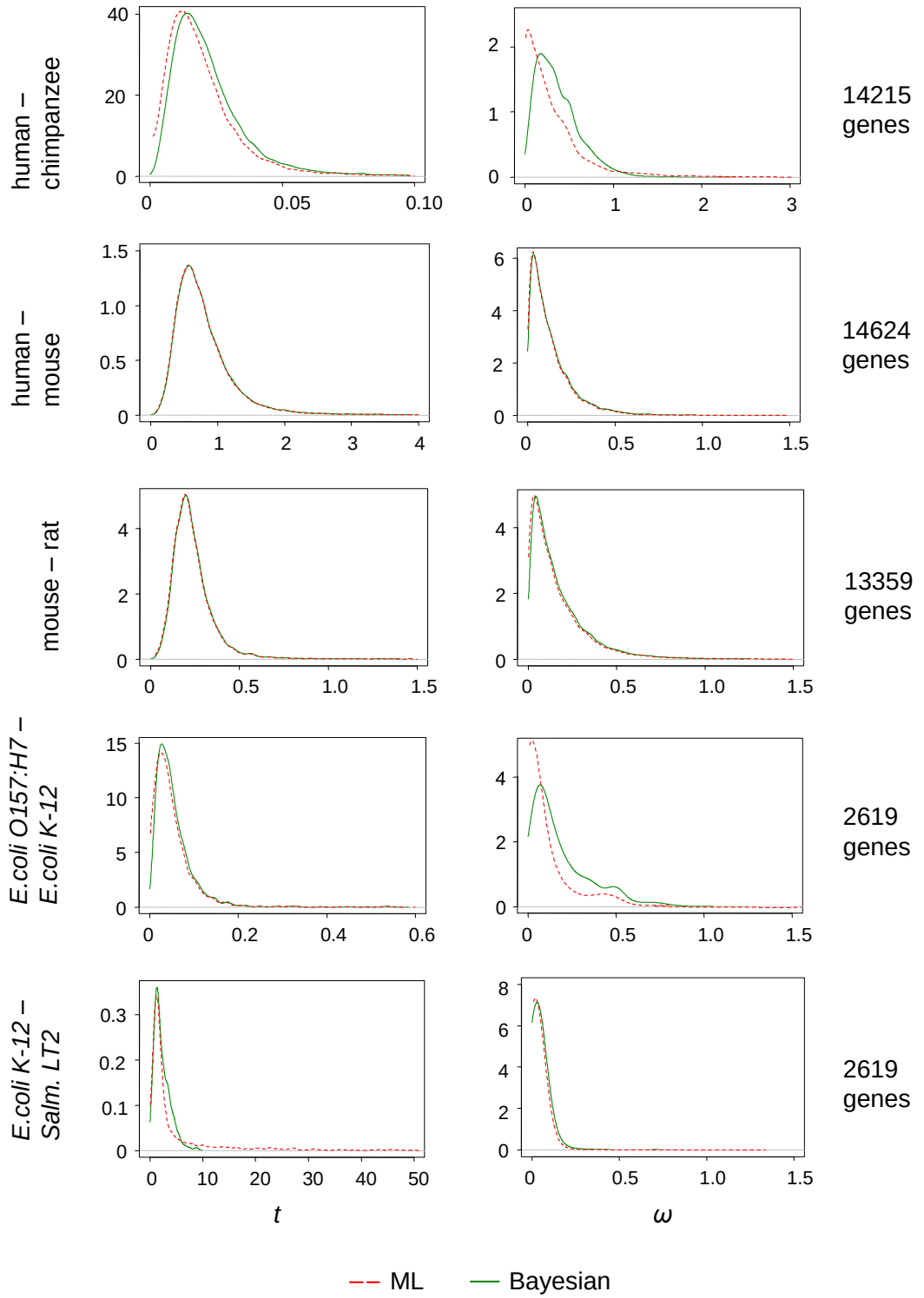

Figure 3.6: Distributions (smoothed histograms) of Bayesian and ML estimates of *t* and *ω* from mammalian and bacterial pairwise gene comparisons. Numbers of genes analyzed in each comparison are reported in the right part of the figure.

Table 3.3: Descriptive statistics of Bayesian (top, bold) and ML (bottom) estimates of *t* and *ω* from pairwise comparisons of protein-coding genes from mammalian species and bacterial strains

		ω							*t*						
				quartiles							quartiles				
	# genes	Mean	SD	25%	50%	75%	N_0	N_∞	Mean	SD	25%	50%	75%	N_0	N_∞
human - chimpanzee	14215	**0.369**	**0.246**	**0.180**	**0.320**	**0.500**	**0**	**0**	**0.025**	**0.072**	**0.013**	**0.019**	**0.028**	**0**	**0**
		0.307	0.418	0.062	0.193	0.411	2507	423	0.022	0.042	0.010	0.016	0.025	377	2
human - mouse	14624	**0.130**	**0.125**	**0.044**	**0.093**	**0.176**	**0**	**0**	**0.812**	**0.574**	**0.503**	**0.691**	**0.958**	**0**	**0**
		0.126	0.157	0.040	0.089	0.170	221	0	0.849	1.252	0.499	0.686	0.952	0	30
mouse - rat	13359	**0.168**	**0.168**	**0.055**	**0.118**	**0.228**	**0**	**0**	**0.242**	**0.179**	**0.163**	**0.215**	**0.281**	**0**	**0**
		0.159	0.180	0.046	0.108	0.215	509	0	0.238	0.232	0.161	0.212	0.278	0	3
E.coli K-12 - *E.coli* O157	2619	**0.179**	**0.170**	**0.055**	**0.116**	**0.252**	**0**	**0**	**0.080**	**0.354**	**0.026**	**0.043**	**0.068**	**0**	**0**
		0.099	0.174	0.001	0.034	0.110	912	31	0.073	0.527	0.020	0.038	0.064	121	6
E.coli K-12 - *Salm.* LT2	2619	**0.037**	**0.042**	**0.016**	**0.025**	**0.042**	**0**	**0**	**2.261**	**1.546**	**1.153**	**1.836**	**3.129**	**0**	**0**
		0.025	0.042	0.006	0.018	0.032	164	0	5.052	8.481	1.087	1.748	4.066	0	217

Note.– Data were analyzed using the F_{61} model for codon frequencies. Results for ML have been calculated after removing the infinite estimates. N_0 is the number of genes with $\hat{\omega}$ or $\hat{t} = 0$, while N_∞ is the number of genes with $\hat{\omega}$ or $\hat{t} = \infty$.

Table 3.4: The numbers of genes with ω estimate greater or less than 1 from pairwise comparisons of protein-coding genes from mammalian species and bacterial strains using the Bayesian and ML methods

Data			Bayesian		
			$\tilde{\omega} < 1$	$\tilde{\omega} > 1$	N_L
human - chimpanzee	ML	$\hat{\omega} < 1$	13094	0	78
		$\hat{\omega} > 1$	822	299	
		N_B		3	
human - mouse	ML	$\hat{\omega} < 1$	14617	0	2
		$\hat{\omega} > 1$	1	6	
		N_B		2	
mouse - rat	ML	$\hat{\omega} < 1$	13313	0	5
		$\hat{\omega} > 1$	10	36	
		N_B		2	
E.coli K-12 - *E.coli* O157	ML	$\hat{\omega} < 1$	2574	0	0
		$\hat{\omega} > 1$	43	2	
		N_B		0	
E.coli K-12 - *Salm.* LT2	ML	$\hat{\omega} < 1$	2617	0	0
		$\hat{\omega} > 1$	2	0	
		N_B		0	

Note.— N_L is the number of genes with statistically significant $\hat{\omega} > 1$ based on the LRT at the 5% level (one-sided with critical value 2.71) in the likelihood method, while N_B is the number of genes with $P(\omega > 1 \mid x) > 0.95$ in the Bayesian analysis.

selection by both methods (Table 3.4). The mouse-rat comparison gives similar patterns to the human-mouse comparison: in both cases, the sequences are moderately divergent and the data are informative.

We re-analyzed the human-chimpanzee and human-mouse alignments using two alternative priors in order to examine the effect of the prior in the posterior estimates of t and ω. The first alternative prior (AP1) is $t \sim G(2, 2)$ and $\omega \sim G(2, 4)$. This has the same means as the default prior of equation (3.3) but the prior here is more informative because of the larger shape parameter (2 vs. 1.1). In the second alternative prior (AP2), we used 2 for the shape parameter, but chose the scale parameter such that the prior mean roughly matches the median of the MLEs for all genes (Table 3.3). Thus for the human-chimpanzee comparison, AP2 is $t \sim G(2, 100)$, with the prior mean 0.02 (while the median of MLEs of t is 0.016), and $\omega \sim G(2, 10)$, with the prior mean 0.2 (while the median of MLEs of ω is 0.193). For the human-mouse comparison, AP2 is $t \sim G(2, 3)$, with the prior mean 0.67 (while the median of the MLEs is 0.686) and $\omega \sim G(2, 20)$, with the prior mean 0.1 (the median of the MLEs is 0.089). In general, it is not advisable to use the data to specify the prior, but in some cases some prior information may be available. For example, between the human and the chimpanzee, the distance t is very likely to be smaller than 0.1.

Posterior estimates of ω and t from the analysis using the default and alternative priors are illustrated in Figure 3.7 and 3.8. In the human-chimpanzee comparison, the impact of the prior is apparent. The Bayesian estimates of ω using the AP1 are higher than those using the default prior for low ω values ($\omega < 0.5$) and lower for high ω values ($\omega > 0.5$) (Figure 3.7A). When the prior is more informative (shape parameter 2), the posterior means are closer to the prior mean 0.5. For the human-mouse comparison estimates under AP1 are close to those under the default prior (Figure 3.7B). The Bayesian estimates of t are less affected by the change in the prior in both comparisons and the estimates are approximately the same for the majority of the genes (Figure 3.8A, B). The effect of the prior AP2 is more significant. In both comparisons the Bayesian estimates of ω are smaller than those obtained using the default prior for almost all genes (Figure 3.7C, D). This is because the priors are more informative (with shape parameter $\alpha = 2$) and have lower means (0.2 and 0.1 for the human-chimpanzee and human-mouse comparisons, respectively, instead of 0.5) and thus affect posterior estimates more than the default prior. The effect is more apparent in the human-chimpanzee comparison because the sequence distances are smaller. Posterior estimates of t are less affected by the change in the prior (Figure 3.8C, D). In summary, the prior affects posterior estimates of ω when the genes are not informative about ω and does not affect significantly the posterior estimates of t.

3.4.2 Analysis of the bacterial data set

We performed two pairwise comparisons: *E. coli* K-12 vs. *E. coli* O157:H7 and *E. coli* K-12 vs. *Salmonella typhimurium* LT2. The two strains of *E. coli* have the same evolutionary distance from the *Salmonella* and gave similar results. Thus, only the results from the *E. coli* K-12 − *Salmonella typhimurium* LT2 comparison are reported here.

The sequences from the two *E. coli* strains are very similar, and the prior has an impact on Bayesian estimates, similar to that in the comparison of the human and chimpanzee genes. The mean, median and 25% and 75% percentiles of the Bayesian ω estimates are 0.179, 0.116 and (0.055, 0.252) while the corresponding results for the MLEs are 0.099, 0.034 and (0.001, 0.110) (Table 3.3). Thus, in the analysis of those genes the two methods give different results. Also, the MLE $\hat{\omega} = 0$ in 912 genes and $\hat{\omega} = \infty$ in 31 genes. None of the genes with $\hat{\omega} > 1$ is statistically significant at the $\alpha = 5\%$ significance level according to the LRT and none has $P(\omega > 1 \mid x) > 0.95$ (Table 3.4). The gene sequences from the *E. coli* K-12 and *Salmonella* are quite divergent. In most genes, the two methods produced similar estimates (Figure 3.6). However, some genes are very divergent with the MLE $\hat{t} = \infty$ in 217 genes.

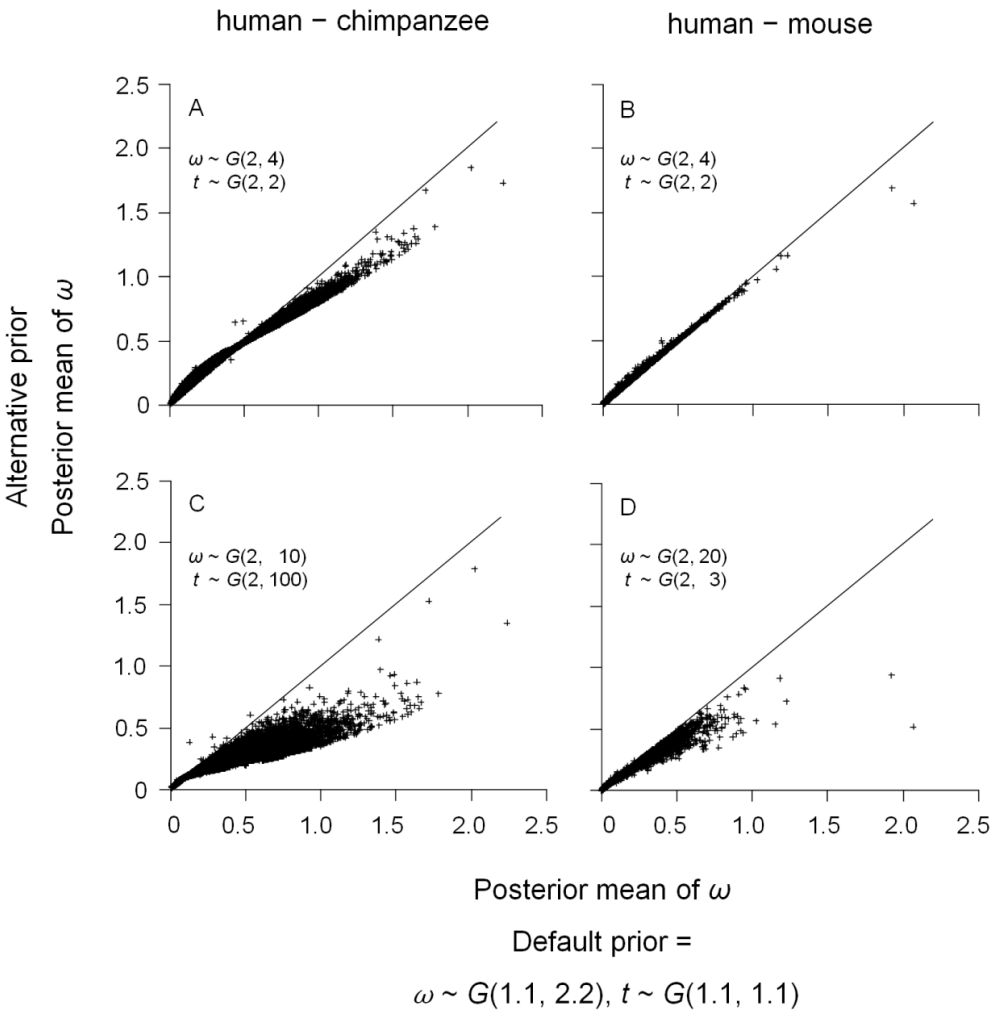

Figure 3.7: Bayesian estimates of ω for the human−chimpanzee (A and C) and human−mouse (B and D) comparisons using alternative priors plotted against estimates using the default prior (equation (3.3)). The alternative priors are: (A and B) $\omega \sim G(2, 4)$, $t \sim G(2, 2)$; (C) $\omega \sim G(2, 10)$, $t \sim G(2, 100)$; and (D) $\omega \sim G(2, 20)$, $t \sim G(2, 3)$.

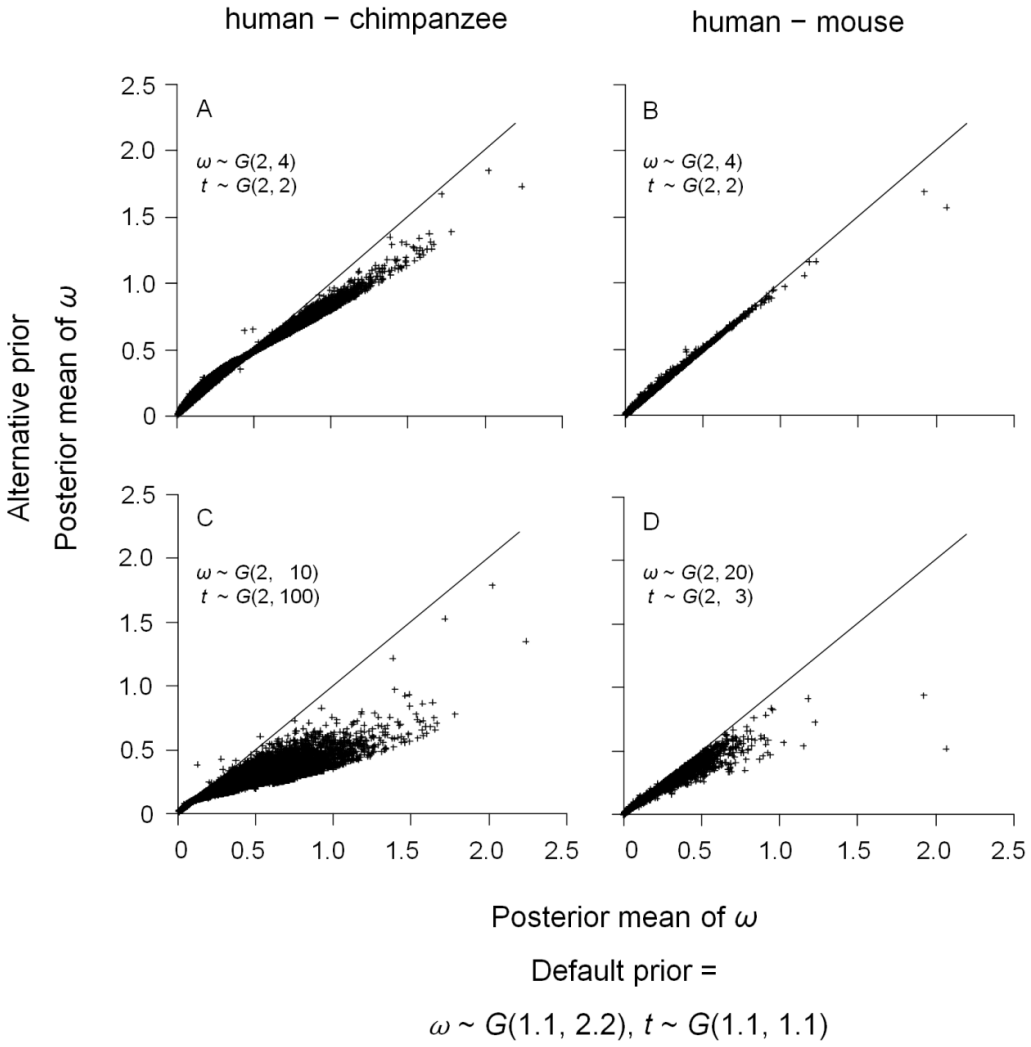

Figure 3.8: Bayesian estimates of *t* for the human−chimpanzee (A and C) and human−mouse (B and D) comparisons using different gamma priors. The alternative priors are as in Figure 3.7.

3.5 Discussion

When sequences from multiple species are available they should be used together in a joint analysis accounting for their phylogenetic relationship. Thus, performing pairwise comparisons to estimate ω is not an efficient use of the data. In particular, a number of likelihood ratio tests have been developed to detect positive selection that affects particular evolutionary lineages on a phylogeny or individual sites in the protein (see for reviews, e.g. Cannarozzi and Schneider 2012; Yang 2014). To apply such tests of positive selection, it is essential to use multiple sequences, as a pair of sequences hardly contains enough

information for the tests to have any power (e.g., Yang 2006). Some proteins may evolve in an episodic manner and thus adaptive episodes may not be detected in pairwise comparisons, especially when the sequences are distantly related (Messier and Stewart 1997). In a pairwise comparison, positive selection is detected only if the ω averaged over all sites in the protein and over the whole evolutionary history connecting the two sequences is higher than one. This seems to be an extremely stringent criterion. Analysis of multiple sequences on a phylogeny allows one to detect episodic positive selection that affects a particular branch (Yang 1998).

Nevertheless, pairwise sequence comparisons are widely used, especially in comparative genomics, sometimes to provide summary statistics of the data and sometimes because of lack of a third genome. The ML method has been widely used to estimate ω and t in pairwise comparisons of genes (e.g., Nielsen, et al. 2005; Ge, et al. 2008; Walters and Harrison 2010; Buschiazzo, et al. 2012; Gladieux, et al. 2013; Wang and Chen 2013). Counting methods are also used due to their simplicity (Garcia-Gil, et al. 2003; Schenekar, et al. 2011; Graves, et al. 2013), even though they were found not to perform as well as ML in computer simulations (Yang and Nielsen 2000). Both counting and ML methods sometimes return 0 or ∞ as estimates, so that neither the expectation nor the variance of the estimates is finite. The infinity estimates of ω appear to be particularly confusing to many users of the methods. For example, some authors added a small arbitrary number (pseudocounts) to the numbers of synonymous and nonsynonymous substitutions before calculating ω to avoid such extreme estimates (e.g. Novaes, et al. 2008; Bajgain, et al. 2011; Pellino, et al. 2013). Other authors excluded genes with $d_S = 0$ from their analyses (e.g., Wang and Chen 2013). The Bayesian method presented here may provide a better procedure than such ad hoc treatments. It always returns finite estimates of ω and t as the prior penalizes extreme values. The computer simulation suggests that the Bayesian estimates of ω have nice statistical properties, with similar or smaller MSEs compared with the MLEs. The posterior means are close to the MLEs when the data are informative, that is, when the sequences are long and the sequence divergence is intermediate, but the differences can be large when the sequences are short and are either too similar or too divergent. Nearly identical sequences contain little information while extremely divergent sequences contain too much noise concerning ω. In both cases, the data are not informative and the prior has an impact on posterior estimates of ω. However, as sequence length increases the effect of the prior decreases irrespective of the true values of ω and t. The Bayesian method described here applies for the analysis of only two sequences. A Bayesian method for the analysis of multiple sequences in a phylogeny requires calculation of high-dimensional integrals and was not pursued.

Note that MLEs $\hat{\omega} = \infty$ should not be taken as evidence for positive selection ($\omega > 1$) because the extreme estimate may well be due to chance effects when the numbers of

changes are small. Instead, positive selection can be claimed only if the LRT is significant in the ML framework or when $P(\omega > 1 \mid x) > 0.95$ in the Bayesian analysis.

The Bayesian method presented here has been implemented in the CODEML program in the PAML package (Yang 2007). The program allows the user to specify the parameters of the gamma priors for t and ω. Although the Bayesian method is computationally more intensive than ML, it remains fast enough for large-scale screening. It takes 1-2 seconds to analyze a pair of sequences on a modern PC.

4 The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times

In the previous chapter we presented a new Bayesian method to estimate the nonsynonymous/synonymous rate ratio for pairwise sequence comparisons and we explored its performance in comparison to ML. In this chapter, we will explore the performance of existing Bayesian inference methods in estimating species divergence times from molecular data (see §2.3 for details) when ancestral polymorphism and incomplete lineage sorting are present in the data. Widely used Bayesian molecular clock dating methods ignore the issue, and it is not clear what impact those aspects of molecular evolution may have on time estimation.

4.1 Introduction

The molecular clock hypothesis states that the rate of evolution of molecular sequences is approximately constant with time (Zuckerkandl and Pauling 1965). This powerful idea means that in practice information from the fossil record can be combined with information from molecular alignments to obtain geological times of divergence for species in a phylogeny. Recently several Bayesian methods have been developed for such type of analysis (Thorne, et al. 1998; Yang 2007; Drummond, et al. 2012; Ronquist, Teslenko, et al. 2012). These methods model important evolutionary processes such as rate variation across lineages (Thorne, et al. 1998; Rannala and Yang 2007) and loci (dos Reis, Zhu, et al. 2014), account for uncertainties in fossil information (Inoue, et al. 2010) and may provide precise time estimates (dos Reis, et al. 2012). Moreover, the development of efficient algorithms has allowed the analysis of large genomic datasets from several species in realistic time frames (Thorne, et al. 1998; dos Reis and Yang 2011). Due to those methodological advances studies of species diversification times using genomic data are nowadays very common (Erwin, et al. 2011; dos Reis, et al. 2012; Jarvis, et al. 2014).

Despite the methodological and computational progress, most molecular-clock dating studies have ignored the effects of the coalescent process on sequence divergences and thus on divergence time estimates. Implicitly, they assume that gene coalescence coincides with

species diversification and gene trees match the species tree; however, this might not always be the case. For example, consider a sample of two nucleotide sequences (genes) belonging to different individuals from a diploid population of N individuals. The expected time to coalescence, that is, the time it takes for the two sequences to find their common ancestor is $2N$ generations (Kingman 1982b, a; Tajima 1983). If the sequences are sampled from individuals belonging to two different, completely isolated species (with no gene flow after speciation) which diverged T generations ago, then the expected sequence coalescent time is $T + 2N$ (Figure 4.1), where N is now the population size of the ancestral species (Gillespie and Langley 1979). In other words, the divergence time of the genes, T^*, can be older than the divergence time of the species (i.e. $T^* > T$), especially so if the size of the ancestral population is large compared to the species divergence time. Note here that the population size N is the effective population size N_e, which is the size of an idealised (Fisher-Wright) population with the same magnitude of genetic drift as the population under study. Such an idealised population is characterised by constant population size, non overlapping generations, random mating and neutral evolution.

Furthermore, for sequences sampled from three or more species, the genealogy of the sequences and the species tree may be in conflict, resulting from the deep coalescent times of the gene sequences (green dashed lines in Figure 4.1), a process known as incomplete lineage sorting (Hudson 1983; Nichols 2001). Thus, studies that use the molecular clock to estimate the times of species divergences from molecular data should take into account the effect of ancestral population size and incomplete lineage sorting on gene ages, otherwise biased estimates of species divergence times may be obtained.

Several Bayesian phylogenetic methods have been developed to perform inference under the multi-species coalescent (Rannala and Yang 2003; Liu and Pearl 2007; Liu 2008; Heled and Drummond 2010; Yang 2015). However, these methods are computationally expensive and are only practical for small datasets or when using simple nucleotide substitution models. Thus, although the coalescent process has long been recognised as an important aspect of sequence evolution (Takahata, et al. 1995; Edwards and Beerli 2000; Kubatko and Degnan 2007; Knowles and Kubatko 2010; Burbrink and Pyron 2011; Oliver 2013; Yang 2014), a majority of molecular clock dating analyses are still carried out ignoring the effects of ancestral population size and incomplete lineage sorting (e.g., Erwin, et al. 2011; dos Reis, et al. 2012; Jarvis, et al. 2014; Misof, et al. 2014; Zeng, et al. 2014). Furthermore, the biases introduced in time estimates by ignoring the coalescent process do not seem to have been studied.

Here, we explore the impact of ancestral polymorphism and incomplete lineage sorting on Bayesian divergence time estimates assuming that the clock holds, when polymorphism and incomplete lineage sorting are ignored by the model. We perform a combination of

mathematical analysis, computer simulations, and analysis of a real dataset (the hominoid phylogeny) and show that ignoring the coalescent process can have a large impact on estimates of divergence times, even when estimating ancient divergence events. Divergence times can be substantially under or overestimated, depending on the configuration and precision of the fossil calibrations on the tree, with the molecular evolutionary rate being usually overestimated. The problem is severe and the results highlight an urgent need for the development of efficient, fast computer software that can provide reliable estimates of divergence times under the multi-species coalescent for the large genome-scale datasets now routinely available.

4.2 The case of three species

Here, we study the case of estimating the two divergence times in a phylogeny of three species when the coalescent process is ignored. We first provide an approximate mathematical formula for the time and rate estimates and their errors when the amount of molecular data (the number of genes or loci) analyzed is very large, when we have perfect fossil information, and when there is little conflict between the species tree and the sampled gene trees. We then use computer simulations to study Bayesian time estimation when incomplete lineage sorting may be substantial and when we use uncertain fossil calibrations in the form of priors.

4.2.1 A simple approximation to the time and rate estimates and their errors when the coalescent process is ignored

Assume the three-species phylogeny of Figure 4.1. We are interested in estimating the two species divergence times on the tree: $t_1 = gT_1$ (the age of the root) and $t_2 = gT_2$ (the age of the internal node), where T is the time in generations, t the time in years, and g the generation time. The probability for a gene tree to be different from the species tree is

$$P = \frac{2}{3} e^{-(T_1 - T_2)/2N} \tag{4.1}$$

(Hudson 1983). Here, we assume that $T_1 - T_2$ is large enough so that the probability of gene tree-species tree mismatch is negligibly small and can be ignored (for example if $T_1 - T_2 > 10N$ then $P < 0.45\%$). In a typical molecular dating analysis, we sample a set of genes from each species, concatenate and align the sequences (i.e. create a supergene alignment) and then estimate the species phylogeny and the molecular distances using the concatenated

alignment. The molecular distances and information from the fossil record are then used to estimate the divergence times. Thus, to understand how time estimates may be affected by ignoring the coalescent process, we must first understand how the molecular distances are affected.

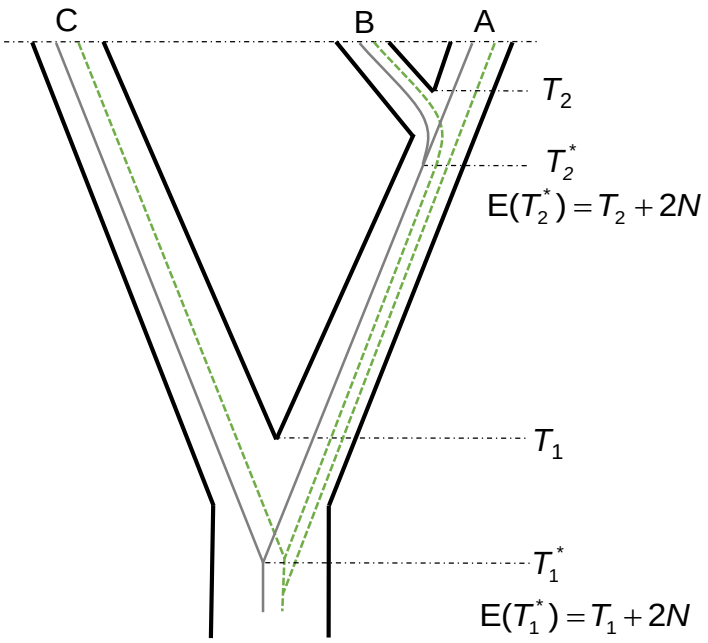

Figure 4.1: A three-species phylogeny. The species tree is represented by thick black lines. The grey lines represent the genealogy for a sample of three genes (one from each species) that matches the species tree. The green dashed lines represent a gene genealogy that does not match the species tree (i.e. we say the species tree and the gene tree are in conflict). If the species have been completely isolated since divergence (i.e. no migration or introgression), then the gene divergence times (T*) will always be older than the species divergence times (T). The expected gene divergence time (in generations) is E(T*) = T + 2N, where N is the size of the ancestral population.

The expected molecular distance (in expected number of substitutions per site) between either of the two genes sampled from A and B and their common ancestor is

$$\mathrm{E}(d_2) = (T_2 + 2N)gr = (t_2 + 2Ng)r, \quad (4.2)$$

where r is the substitution rate per site per year (so that $\mu = rg$ is the rate per generation), which we assume to be the same for all loci in all lineages. Similarly, for a sample of two genes from A (or B) and C, the expected distance is

$$\mathrm{E}(d_1) = (T_1 + 2N)gr = (t_1 + 2Ng)r. \quad (4.3)$$

Note that equations (4.2) and (4.3) are the expected distances for a pairwise sequence alignment for a single locus. If our supergene alignment is very long (so that it contains a

large number of loci) the molecular distances (the branch lengths) estimated on the species tree will be close to the expected values

$$\hat{d}_2 \approx \mathrm{E}(d_2) = (t_2 + 2Ng)r\,, \tag{4.4}$$

and

$$\hat{d}_1 \approx \mathrm{E}(d_1) = (t_1 + 2Ng)r\,. \tag{4.5}$$

However, as we sample more and more loci, the distance estimates will not converge to the expectations in equations (4.2) and (4.3) because they are estimated on the species tree (and not on the pairwise alignments) and incomplete lineage sorting is ignored. Nevertheless, we show below that the approximations are quite good even when there is substantial incomplete lineage sorting.

Now let's assume that the age of the root, t_1, is known (say, from the fossil record). Then under the molecular clock, an estimator of t_2 using t_1 as a calibration is

$$\hat{t}_2 = t_1\hat{d}_2 \,/\, \hat{d}_1. \tag{4.6}$$

The estimator is constructed under the molecular clock hypothesis that the ratio of the species divergence times (t_2/t_1) is the same as the ratio of the molecular distances (d_2/d_1). However, the later ratio is the ratio of gene divergences, and thus the estimator of the species A and B divergence time $\hat{t}_2$ will be biased. If we replace the distance estimates in equation (4.6) with their approximations from equations (4.4) and (4.5) we obtain an approximation for $\hat{t}_2$ as a function of the true parameter values

$$\hat{t}_2 \approx t_1 \frac{(t_2 + 2Ng)}{(t_1 + 2Ng)}\,. \tag{4.7}$$

Then, an approximation to the bias of that estimator is

$$\hat{t}_2 - t_2 \approx t_1 \frac{(t_2 + 2Ng)}{(t_1 + 2Ng)} - t_2 = \frac{(t_1 - t_2)2Ng}{t_1 + 2Ng}\,. \tag{4.8}$$

Because $t_1 > t_2$ the bias is positive and thus t_2 is overestimated. In contrast, if the age of the internal node, t_2, is known (i.e. from the fossil record), we can instead estimate the age of the root as

$$\hat{t}_1 = t_2 \frac{\hat{d}_1}{\hat{d}_2} \approx t_2 \frac{(t_1 + 2Ng)}{(t_2 + 2Ng)}\,, \tag{4.9}$$

with approximate bias

$$\hat{t}_1 - t_1 \approx \frac{(t_2 - t_1)2Ng}{t_2 + 2Ng}\,. \tag{4.10}$$

Here, the bias is always negative and t_1 is underestimated.

Similarly, we can estimate the molecular rate, r, using t_1 as the calibration time

$$\hat{r} = \hat{d}_1 \,/\, t_1\,. \tag{4.11}$$

If we replace the distance estimate with its approximation from equation (4.5) we get

$$\hat{r} \approx \frac{(t_1 + 2Ng)r}{t_1} = r + \frac{2Ngr}{t_1}. \quad (4.12)$$

Thus, the approximate bias of the rate estimator is

$$\hat{r} - r \approx 2Ngr / t_1. \quad (4.13)$$

Here, the bias is always positive and thus r is overestimated. We can also estimate the rate in a similar way when t_2 is used as the calibration time

$$\hat{r}' = \hat{d}_2 / t_2 \approx r + 2Ngr / t_2, \quad (4.14)$$

with approximate bias

$$\hat{r}' - r \approx 2Ngr / t_2. \quad (4.15)$$

Here, the bias is also positive and thus the rate is overestimated no matter whether t_1 or t_2 is used as calibration. However, since $t_1 > t_2$ the overestimation is more severe when t_2 is used as the calibration time.

The relative error of an estimator ($\hat{\theta}$) is the bias of the estimator divided by the true parameter value (θ)

$$\varepsilon\left(\hat{\theta}\right) = \left(\hat{\theta} - \theta\right) / \theta. \quad (4.16)$$

Thus, we can use the biases of equations (4.8), (4.10), (4.12) and (4.15) to obtain approximations to the relative errors on the estimates of t_1, t_2 and r. Note that if the relative error is positive, then the parameter is overestimated, and if it is negative the parameter is underestimated. Figure 4.2 and 4.3 show the relative errors on estimates of t_1, t_2 and r for a few cases when the coalescent process is ignored. In some cases the errors can be substantial. For example, when t_2 = 1 Ma, t_1 = 10 Ma, g = 10 years (y) and $N = 10^5$ individuals, t_2 is overestimated by 150% when using t_1 as the calibration time (Figure 4.2A). On the other hand, for the same parameter values and when t_2 is used as the calibration time, t_1 is underestimated by 60% (Figure 4.2B) and r is overestimated by 200% (Figure 4.3). Note that for those parameter values the species tree-gene tree mismatch is very small (P = 0.74%).

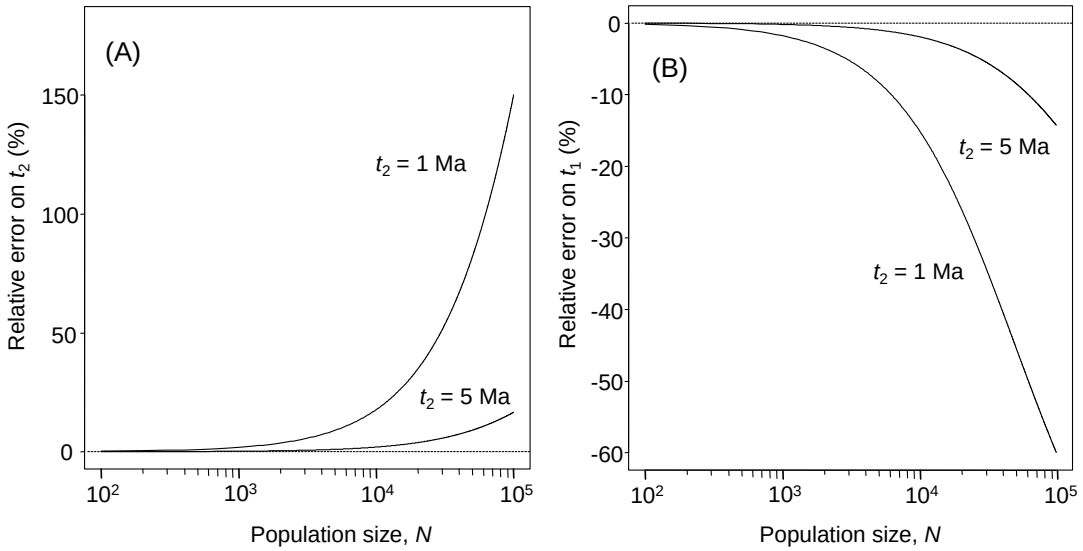

Figure 4.2: Relative errors in estimates of divergence times on a three-species phylogeny (Figure 4.1) as a function of population size when the coalescent process is ignored. The errors are calculated approximately using equations (4.8), (4.10) and (4.16). (A) Relative errors of estimates of the internal node's age, t_2, when the age of the root, t_1, is known and used as the calibration. (B) Relative errors of t_1 estimates when t_2 is the calibration. In (A) and (B) the true values are t_1 = 10 Ma, t_2 = 1 or = 5 Ma, and g = 10 y.

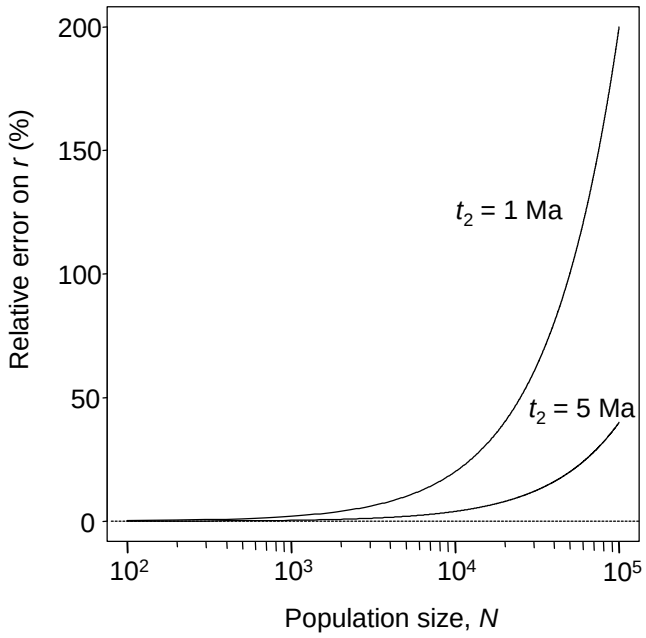

Figure 4.3: Relative errors in estimates of the molecular substitution rate on a three-species phylogeny (Figure 4.1) as a function of population size when the coalescent process is ignored. The error is calculated approximately using equations (4.15) and (4.16), with t_2 known and used as the calibration. The true values are t_2 = 1 or = 5 Ma, g = 10 y, and r = 10^{-9} s/s/y.

4.2.2 Simulation analysis: Bayesian estimates of times when the coalescent process is ignored

We simulated 50 gene alignments (each alignment of 1,000 nucleotides) from a three-species phylogeny (Figure 4.1) using the program MCCOAL (Rannala and Yang 2003; Yang and Rannala 2010). MCCOAL simulates gene trees under the multi-species coalescent with corresponding gene alignments using the JC69 substitution model. Thus the simulated gene trees may not match the species tree. The species divergence times are t_1 = 10 Ma, and t_2 = 1, 5, 9 Ma. The generation time is g = 10 years, and the substitution rate is $r = 10^{-9}$ s/s/y. The population size (assumed to be constant in all lineages) is $N = 10^2, 10^3, 10^4, 10^5, 10^6$ individuals. This gives a total of 3×5 = 15 parameter combinations. The number of replicates (the number of times each parameter setup is simulated) is 100.

We concatenated the simulated gene alignments into a supergene alignment, and we obtained Bayesian estimates of divergence times under the clock and the JC69 model using the program MCMCTREE (Yang 2007). Note that the MCMCTREE does not account for ancestral polymorphism and incomplete lineage sorting. The species tree (Figure 4.1) is used and assumed known. The time unit was set at 10 My. We used a diffuse gamma prior for the rate, $r \sim G(1, 100)$, with mean 0.01 per time unit (i.e. meaning 10^{-9} s/s/y).We used two different strategies to construct the time prior. *Strategy 1*: The prior on the age of the root is $t_1 \sim G(100, 100)$. This is an informative prior, equivalent to a fossil calibration with mean 10 Ma and 95% prior interval 8–12 Ma. For t_2 we used a diffuse prior density conditioned on t_1 (a uniform distribution between 0 and t_1). *Strategy 2*: We used informative calibrations on both times, $t_1 \sim$ B(0.7, 1.4) and $t_2 \sim$ B(0.4, 0.6), equivalent to 7–14 My and 4–6 Ma respectively. Here B(t_L, t_U) means that the time is uniformly distributed between a minimum age t_L and a maximum age t_U, with 5% probability that the time is outside the interval (2.5% on each side). Note the calibration on the root is more uncertain than that on t_1 (with the uncertainty measured by the calibration width divided by the midpoint of the calibration, as in dos Reis and Yang 2013a). The first strategy reflects good fossil information for the root and absence of fossil information for the internal node while the second represents good fossil information for the internal node and uncertain for the root. The simulated data, D, were analysed under both calibration strategies, and the posterior mean of the times and rate, $\tilde{t} = E(t \mid D)$ and $\tilde{r} = E(r \mid D)$, their relative errors, and the 95% credibility intervals (CIs) were collected and averaged among the 100 replicates. For each replicate the MCMC algorithm was run with a burn-in of 5×10^6 iterations collecting 10,000 samples from the posterior every 2,000 iterations.

Table 4.1 shows a few summary statistics for the simulated data sets. The amount of incomplete lineage sorting (i.e. the probability of conflict between gene trees and the species tree) varied from 0% (for t_2 = 1 Ma, t_1 = 10 Ma and $N = 10^2$) up to 63.4% (for t_2 = 9 Ma, t_1 = 10 Ma and $N = 10^6$). Table 4.1 also shows the MLEs of the molecular distances for the supergene alignment obtained on the species tree using the program BASEML (Yang 2007). The estimated distances are virtually identical to the expectations (equations 4.2 and 4.3) when incomplete lineage sorting is negligible; and they are still very close to the expectations even when incomplete lineage sorting is substantial (bold lines in Table 4.1). This shows that the long supergene alignment is very informative about the molecular distances (i.e. there is little error in the MLE of the distances) and that the approximations of the equations (4.4) and (4.5) are very good even with substantial incomplete lineage sorting.

Table 4.1: Estimates of divergence times and their errors as a function of population size in a three-species phylogeny.

t_2	N	P	E(d_2)	$\hat{d}_2$	E(d_1)	$\hat{d}_1$	$\hat{t}_2$	$\varepsilon(\hat{t}_2)$	$\hat{t}_1$	$\varepsilon(\hat{t}_1)$
1	10^2	0.000	0.0010	0.0010	0.0100	0.0100	1.00	0%	9.98	–0.2%
	10^3	0.000	0.0010	0.0010	0.0100	0.0100	1.02	2%	9.82	–1.8%
	10^4	0.000	0.0012	0.0012	0.0102	0.0102	1.18	18%	8.50	–15.0%
	10^5	0.007	0.0030	0.0030	0.0120	0.0120	2.50	150%	4.00	–60.0%
	10^6	**0.425**	**0.0210**	**0.0211**	**0.0300**	**0.0297**	**7.00**	**600%**	**1.43**	**–85.7%**
5	10^2	0.000	0.0050	0.0050	0.0100	0.0100	5.00	0%	10.00	0%
	10^3	0.000	0.0050	0.0050	0.0100	0.0100	5.01	0.2%	9.98	–0.2%
	10^4	0.000	0.0052	0.0052	0.0102	0.0102	5.10	2.0%	9.81	–1.9%
	10^5	0.055	0.0070	0.0070	0.0120	0.0120	5.83	16.6%	8.57	–14.3%
	10^6	**0.519**	**0.0250**	**0.0243**	**0.0300**	**0.0294**	**8.33**	**66.6%**	**6.00**	**–40%**
9	10^2	0.000	0.0090	0.0090	0.0100	0.0100	9.00	0%	10.00	0%
	10^3	0.000	0.0090	0.0090	0.0100	0.0100	9.00	0%	10.00	0%
	10^4	0.005	0.0092	0.0092	0.0102	0.0102	9.02	0.2%	9.98	–0.2%
	10^5	0.404	0.0110	0.0110	0.0120	0.0120	9.17	1.9%	9.82	–1.8%
	10^6	**0.634**	**0.0290**	**0.0280**	**0.0300**	**0.0295**	**9.67**	**7.4%**	**9.31**	**–6.9%**

Note.− Times are in My. The age of the root is t_1 = 10 Ma, and the generation time is g = 10 y. The time estimates are calculated using equations (4.7) and (4.9), and the relative errors with equations (4.8) and (4.10). $P = 2/3 \exp[-(T_1 - T_2)/(2N)]$ is the species tree-gene tree mismatch probability. E(d_2) and E(d_1) are the expected molecular distances from the tips of the phylogeny to the respective coalescent events (equations (4.2) and (4.3)). The molecular distance estimates, $\hat{d}_2$ and $\hat{d}_1$, are obtained from data simulated with the program MCCOAL and estimated by maximum likelihood using the program BASEML, under the clock, on the species phylogeny, and averaged over the 100 replicates.

Figure 4.4 and Table 4.2 show Bayesian estimates of times and of the molecular rate as a function of the population size for the simulated data. Under calibration strategy 1, the age of the root is accurately estimated in all cases owing to the informative calibration on t_1. On the other hand, t_2 is overestimated, with the estimate's error becoming increasingly worse with increasing population size (Figure 4.4A-C). For example, for $N = 10^6$ and $t_2 = 1$ Ma, $\tilde{t}_2$ is 7.28 Ma, i.e. a relative error of 628% (Figure 4.4A). The rate is also overestimated as N increases, irrespective of the true age of the internal node. For example, for $N = 10^6$, $\tilde{r} = 2.91\times10^{-9}$ s/s/y (relative error 191%) for $t_2 = 5$ Ma (Figure 4.4B', Table 4.2). In calibration strategy 2, t_2 has the most precise (or informative) calibration, and so this calibration dominates the analysis. The age of the root in this case is underestimated (as expected from equation 4.10), and the rate is overestimated, as N increases. Also, the rate is more significantly overestimated than in calibration strategy 1 (relative error 317% vs. 191% for $N = 10^6$, Table 4.2) as expected according to equation (4.15).

The posterior time and rate estimates in Figure 4.4 are close to the approximations for the estimators (solid line) calculated with equations (4.7), (4.9), (4.12) and (4.14). Note that the estimates of times and rate were derived without reference to any particular nucleotide substitution model. Thus the theory of equations (4.7), (4.9), (4.12) and (4.14) is also expected to apply to simulations carried out under more complex substitution models such as HKY or GTR (Hasegawa, et al. 1985; Yang 1994a), that is, we expect to see the same biases and relative errors on the estimates. The use of JC69 in our simulations here is thus unimportant, and has no effect on the properties of sequence evolution under the multi-species coalescent.

4.2.3 Simulation analysis: Bayesian estimates of times under the multi-species coalescent

We re-analyzed the simulated gene alignments on the three-species phylogeny with the program BPP (Yang 2015), which can be used to obtain estimates of relative divergence times among species, τ, under the multi-species coalescent. It thus accounts for ancestral polymorphism and incomplete lineage sorting. The relative times are given as expected number of substitutions per site (i.e. they are the molecular distances between the tips of the phylogeny and the species divergence events, so that $\tau = rt$), and so we devised a method to translate these relative time estimates into geological times. In this section we aim to highlight how analysis under the correct model (the multi-species coalescent) can produce time estimates that are unbiased and have little error.

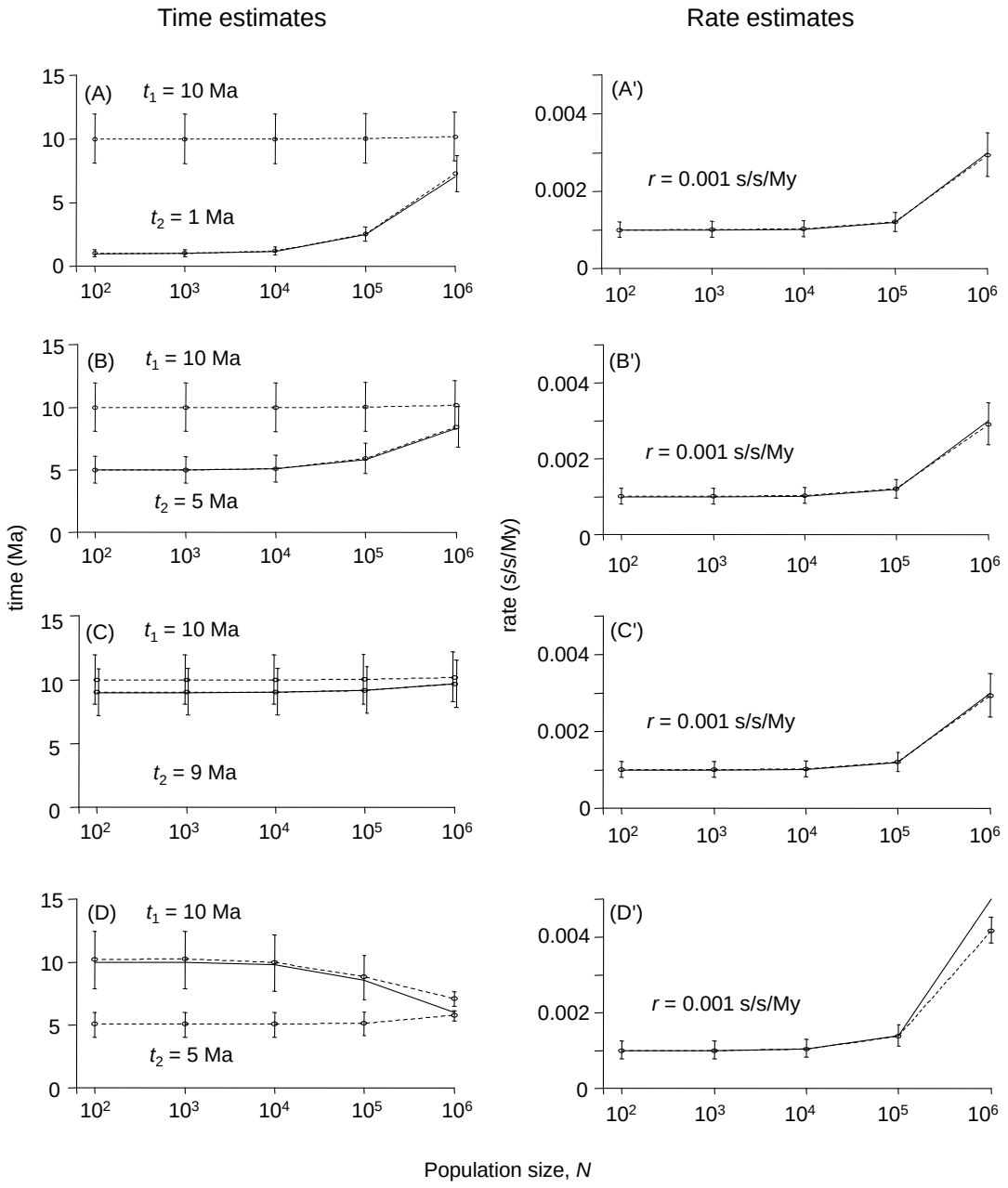

Figure 4.4: Bayesian estimates of divergence times (A-D) and the molecular rate (A'-D') for simulated data on a three-species phylogeny. The data were simulated under the multi-species coalescent, but the coalescent process is ignored during Bayesian estimation of divergence times with the program MCMCTREE. In all cases the true rate is r = 0.001 s/s/My. In (A-C) and (A'-C') the root has the most precise calibration, $t_1 \sim G(100, 100)$ while the internal node has a diffuse prior density, $t_2|t_1 \sim U(0, t_1)$. In these cases the age of the root is correctly estimated, but the age of the internal node and the molecular rate are both progressively overestimated with larger N. In (D, D') the internal node has the most precise calibration, $t_2 \sim B(0.4, 0.6)$ vs. $t_1 \sim B(0.7, 1.4)$. In this case the age of the root is progressively underestimated, and the molecular rate is overestimated, with larger N. The solid lines indicate estimates for t_2 (in A-C) or t_1 (in D) and r (in A'-D') calculated using the estimators of equations (4.7), (4.9) and (4.12), (4.14), respectively.

Table 4.2: Posterior means, 95% CIs, and relative errors of divergence times estimates (in My) and molecular rate for a three-species phylogeny.

Software/Calibrations	*N*	$\tilde{t}_1$	(95% CI)	$\varepsilon(\tilde{t}_1)$	$\tilde{t}_2$	(95% CI)	$\varepsilon(\tilde{t}_2)$	$\tilde{r}$ (× 10^{-3})	(95% CI)	$\varepsilon(\tilde{r})$
MCMCTREE	10^3	10.00	(8.10, 11.97)	0.0%	4.99	(3.96, 6.06)	–0.2%	1.01	(0.81, 1.23)	1%
$t_1 \sim G(100, 100)$	10^4	10.00	(8.09, 11.97)	0.0%	5.11	(4.05, 6.21)	2.2%	1.03	(0.83, 1.24)	3%
$t_2 \mid t_1 \sim U(0, t_1)$	10^5	10.02	(8.11, 11.99)	0.2%	5.90	(4.71, 7.14)	18.0%	1.21	(0.97, 1.45)	21%
	10^6	10.19	(8.28, 12.16)	1.9%	8.43	(6.82, 10.09)	68.6%	2.91	(2.36, 3.49)	191%
MCMCTREE	10^3	10.25	(7.89, 12.47)	2.5%	5.08	(4.03, 6.02)	1.6%	1.00	(0.79, 1.25)	0%
$t_1 \sim B(0.7, 1.4)$	10^4	10.01	(7.72, 12.16)	0.1%	5.08	(4.03, 6.02)	1.6%	1.04	(0.82, 1.30)	4%
$t_2 \sim B(0.4, 0.6)$	10^5	8.84	(7.02, 10.51)	–11.6%	5.16	(4.17, 6.04)	3.2%	1.38	(1.12, 1.69)	38%
	10^6	7.10	(6.51, 7.64)	–29.0%	5.78	(5.34, 6.15)	15.6%	4.17	(3.85, 4.53)	317%
BPP	10^3	10.00	(8.08, 11.98)	0.0%	4.98	(3.94, 6.06)	–0.4%	1.01	(0.81, 1.22)	1%
$t_1 \sim G(100, 100)$	10^4	10.00	(8.09, 11.98)	0.0%	5.02	(3.93, 6.15)	0.4%	1.01	(0.81, 1.22)	1%
	10^5	10.00	(8.07, 11.97)	0.0%	5.10	(3.67, 6.61)	2.0%	1.02	(0.80, 1.25)	2%
	10^6	10.00	(8.09, 11.98)	0.0%	5.23	(2.43, 8.12)	4.6%	1.00	(0.75, 1.27)	0%

Note.− The true values are $t_1 = 10$ Ma, $t_2 = 5$ Ma and $r = 10^{-3}$ s/s/My. Posterior means and 95% CIs are averaged across 100 replicate analyses. The τ (distance) estimates from BPP were translated into absolute geological times by sampling from $t_1 \sim G(100, 100)$.

4.2.4 Simulation analysis: Bayesian estimates of times under the multi-species coalescent

We re-analyzed the simulated gene alignments on the three-species phylogeny with the program BPP (Yang 2015), which can be used to obtain estimates of relative divergence times among species, τ, under the multi-species coalescent. It thus accounts for ancestral polymorphism and incomplete lineage sorting. The relative times are given as expected number of substitutions per site (i.e. they are the molecular distances between the tips of the phylogeny and the species divergence events, so that $\tau = rt$), and so we devised a method to translate these relative time estimates into geological times. In this section we aim to highlight how analysis under the correct model (the multi-species coalescent) can produce time estimates that are unbiased and have little error.

Note that in BPP, the gene alignments *are not* concatenated in a single alignment. Sequences are analysed under the JC69 model and under the clock. We assigned a gamma prior on the relative age of the root, $\tau_1 \sim G(2, 200)$, with mean 0.01 (the true value of τ_1). For τ_2 we used a diffuse prior conditioned on τ_1 (uniform between 0 and τ_1, see Yang and Rannala 2010, equation 2). For the population size parameters, $\theta = 4N\mu$, we used a gamma prior, $\theta \sim G(2, \beta)$, with β set so that the mean of the distribution matches the true population size in the simulations. BPP estimates one θ per ancestral lineage (i.e. two values for the phylogeny of Figure 4.1, one for the AB ancestral lineage, and another for the ABC lineage beyond the root). The same mutation rate was assumed across all loci. For each replicate the MCMC algorithm was run with burn-in 200,000 iterations collecting 10,000 samples from the posterior sampling every 100 iterations.

The relative divergence times estimated with BPP are not directly comparable to the times estimated with MCMCTREE and thus we translated them into absolute geological times by using either a fossil calibration or a prior on the per year mutation rate, r. We used the following procedure. Consider an MCMC sample from the posterior distribution of relative ages (i.e., the i-th sample of the relative root age is τ_1^i) obtained with BPP. First, we sampled values, t_1^i, from a prior density on the root age $t_1 \sim G(100, 100)$. Then samples for the age of the internal node and the per year mutation rate are given by $t_2^i = t_1^i \tau_2^i / \tau_1^i$ and $r_i = \tau_1^i / t_1^i$, respectively. We simply sampled as many values of t_1^i as the number of samples in the MCMC. In this way we obtain a posterior sample of t_1, t_2 and r under the multi-species coalescent (the posterior of t_1 is simply the prior sampling density). The resulting sample can

be summarised in the usual way to obtain the posterior mean of times, rate and 95% CIs. Averages of posterior means and 95% CIs across the 100 replicates are reported on Table 4.2.

Divergence times estimated with BPP for the case t_2 = 5 Ma are shown on Table 4.2. The posterior means for t_2 and r are very accurate (close to the true values) with little relative error. Furthermore, the 95% CIs always contain the true values, even when the mismatch probability between gene trees and the species tree is high. However, for large population sizes the uncertainty around $\tilde{t}_2$ can be quite large because of substantial variation in the coalescent times across genes. For example, for $N = 10^6$ the CI is 2.43−8.12 Ma. Estimates for the cases t_2 = 1 and t_2 = 9 Ma show similar patterns (high accuracy and low error) and are not reported here.

4.3 The case of nine-species

In the case of a three-species phylogeny we saw that the molecular rate is overestimated when the coalescent process is ignored and that time estimates may be under or over estimated depending on which node has the most precise fossil calibration. For larger phylogenies with multiple fossil calibrations the situation is more complicated. Here, we use computer simulation and Bayesian analysis to explore the effect of ancestral polymorphism and incomplete lineage sorting on estimation of divergence times on a phylogeny of nine species with multiple fossil calibrations.

We simulated gene samples for 50 loci (each 1,000 nucleotides long) on the nine-species phylogeny of Figure 4.5 using the MCCOAL program. We considered two cases: (1) a young phylogeny where the root is 10 Ma; and (2) an old phylogeny where the root is 100 Ma. The true ages of the internal nodes are shown in Figure 4.5. In both cases the true substitution rate is $r = 10^{-9}$ s/s/y, and the generation time is g = 10 y. We used two scenarios for the population size. In the first, we simulated gene alignments on the two phylogenies assuming a constant population size in all lineages, with $N = 10^3, 10^4, 10^5, 10^6$ individuals. In the second, we simulated a more realistic case where N varied among lineages (between 10^3 and 10^6 individuals, Figure 4.5). In total we simulated 10 cases (2 phylogenies × 5 population size cases). The true times and rate were used to calculate the relative ages, τ, needed by the program MCCOAL in the simulation. For example, in the young phylogeny the relative age of the root is τ_{10} = 10 My × 0.001 s/s/My = 0.01 s/s, while in the old phylogeny it is τ_{10} = 100 My × 0.001 s/s/My = 0.1 s/s. The population sizes were translated into population size parameters (θs) as well, through $\theta = 4N\mu$. The number of simulation replicates was 100.

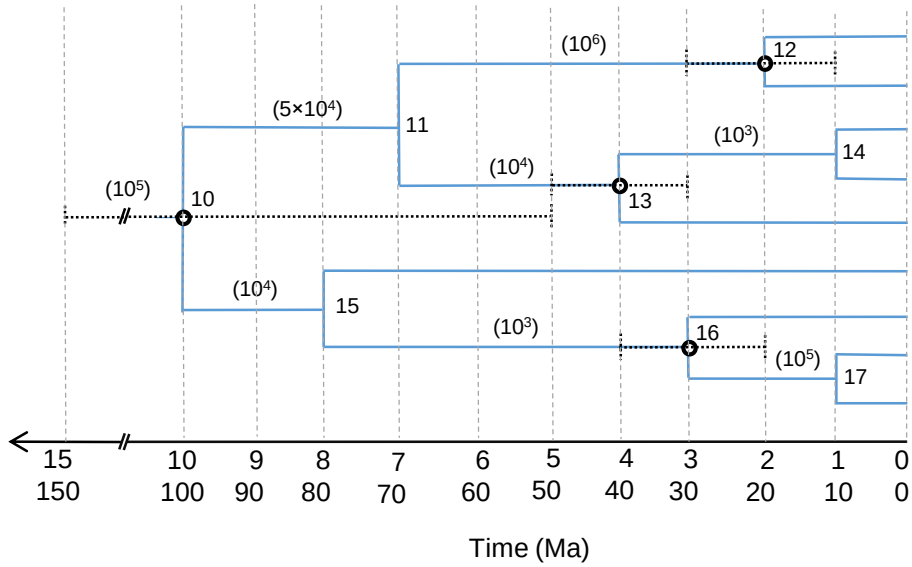

Figure 4.5: A nine-species phylogeny used to simulate gene alignments under the multi-species coalescent. The fossil constraints used for Bayesian estimation of divergence times with the program MCMCTREE are shown as dotted bars. The numbers in brackets correspond to the ancestral population sizes, *N*, for the corresponding branches for the case of variable *N* among lineages. The nodes are numbered from 10 to 17.

The simulated alignments were concatenated into a supergene alignment and analyzed with the MCMCTREE program to estimate the species divergence times and the rate. Analyses were carried out under the clock and under the JC69 model of nucleotide substitution. The parameters of the birth-death model with species sampling used to specify the time prior on nodes without fossil calibrations were set to $\lambda = \mu = 1$ and $\rho = 0$ (Yang and Rannala 2006). These values specify a diffuse uniform kernel density on the node ages. The time unit was set to be 10 My for the young phylogeny and 100My for the old one. We used diffuse priors on the rate: $r \sim G(1, 100)$ and $r \sim G(1, 10)$ for the young and old phylogenies, respectively, with prior means 0.01 and 0.1 substitutions per time unit, respectively, with both meaning 10^{-9} s/s/y. Four nodes have soft fossil calibrations: $t_{10} \sim$ B(0.5, 1.5), $t_{12} \sim$ B(0.1, 0.3), $t_{13} \sim$ B(0.3, 0.5) and $t_{16} \sim$ B(0.2, 0.4), where, for example, B(0.5, 1.5) means that the divergence time is between 5 and 15 Ma in the young phylogeny, or between 50 and 150 Ma in the old phylogeny (Yang and Rannala 2006). The posterior mean of the times and rate, their relative errors, and the 95% CIs were collected and averaged among the 100 replicates.

Table 4.3 shows the posterior means of times and rate and their relative errors averaged across all replicates. For the young phylogeny, when *N* is small (10^3 and 10^4) there is no incomplete lineage sorting (*P* = 0 %) and node ages are overestimated, with relative errors

ranging from 5% to 20% (Table 4.3). As N increases (10^5, 10^6) the ages of nodes close to the root (t_{10}, t_{11}, t_{13}, t_{15}) become increasingly underestimated, while the ages of the external nodes (t_{12}, t_{14}, t_{16}, t_{17}) become increasingly overestimated. For the larger N values the amount of incomplete lineage sorting is substantial and the relative errors in time and rate estimates can be quite dramatic. For example, for $N = 10^6$, the age of the root is underestimated from 10 Ma to 4.8 Ma (–52% error) while the age of a young node (17) is overestimated from 1 Ma to 2.6 Ma (160%), and the molecular rate is overestimated by 628% (Table 4.3). Note that the time estimates of all nodes are concentrated between 2.6 Ma and 4.8 Ma. This is because the most informative calibrations on nodes 12, 13, and 16 range from 1 Ma to 5 Ma. Similar trends can be noticed when N varies among lineages. For example, the ages for external nodes with large ancestral population sizes (i.e. nodes 12 and 17) were overestimated (errors 105% and 50%, respectively), while the ages for external nodes with small ancestral population sizes (i.e. nodes 14, 16) were underestimated (errors −30% and −33%, respectively), as did the ages for the nodes close to the root (i.e. nodes 10, 11, 15) (Table 4.3).

Results for the old phylogeny were similar to the young phylogeny, although the errors in the estimates are smaller (Table 4.3). This is because in the old phylogeny there is substantially less incomplete lineage sorting, and the discrepancies between gene divergence times and species divergence times are less severe. For example, for $N = 10^6$ and $g = 10$ y we expect genes to coalesce at $2Ng = 20$ My over the speciation event, so genes that enter the ancestral population at the root of the phylogeny, would have an expected divergence time of 120 Ma in the old phylogeny, or 20% older than the root speciation event at 100 My. However, for the small phylogeny, the equivalent case means an expected gene divergence age of 30 Ma, or 200% older than the root speciation event at 10 My. This conflict between gene ages and species ages clearly leads to the errors in the divergence time estimates. Note that although the situation is not as severe in the old phylogeny, the relative errors are still substantial. For example, in the old phylogeny, for $N = 10^6$, the relative error on the age of the root is –20.5%, while for one of the younger nodes (node 14) the error is 93%, and for the molecular rate the error is 54% (Table 4.3). For both the young and old phylogenies, the molecular rate is overestimated when the amount of incomplete lineage sorting is substantial (Table 4.3).

Table 4.3: Posterior means of divergence times and molecular rate and their relative errors for the nine species phylogenies for various population sizes.

	N	$\tilde{t}_{10}$ (error)	$\tilde{t}_{11}$ (error)	$\tilde{t}_{12}$ (error)	$\tilde{t}_{13}$ (error)	$\tilde{t}_{14}$ (error)	$\tilde{t}_{15}$ (error)	$\tilde{t}_{16}$ (error)	$\tilde{t}_{17}$ (error)	$\tilde{r}$ (error)	*P* (%)
Young Phylogeny		t_{10} **= 10**	t_{11} **= 7**	t_{12} **= 2**	t_{13} **= 4**	t_{14} **= 1**	t_{15} **= 8**	t_{16} **= 3**	t_{17} **= 1**	***r* = 1**	
	10^3	10.6 (6.0%)	7.4 (5.7%)	2.1 (5.0%)	4.2 (5.0%)	1.1 (10.0%)	8.5 (6.3%)	3.2 (6.7%)	1.1 (10.0%)	0.97 (–3.0%)	0
	10^4	10.1 (1.0%)	7.2 (2.9%)	2.2 (10.0%)	4.2 (5.0%)	1.2 (20.0%)	8.2 (2.5%)	3.2 (6.7%)	1.2 (20.0%)	1.02 (2.0%)	0
	10^5	8.0 (–20.0%)	6.0 (-14.3%)	2.5 (25.0%)	4.0 (0.0%)	1.9 (90.0%)	6.6 (-17.5%)	3.3 (10.0%)	1.9 (90.0%)	1.55 (55.0%)	69
	10^6	4.8 (–52.0%)	4.3 (-38.6%)	2.9 (45.0%)	3.6 (-10.0%)	2.7 (170.0%)	4.3 (-46.3%)	3.4 (13.3%)	2.6 (160.0%)	7.28 (628.0%)	100
	Variable	7.4 (–26.0%)	4.9 (-30.0%)	4.1 (105.0%)	2.8 (-30.0%)	0.7 (-30.0%)	5.1 (-36.3%)	2.0 (-33.3%)	1.5 (50.0%)	1.62 (62.0%)	64
Old Phylogeny		t_{10} **= 100**	t_{11} **= 70**	t_{12} **= 20**	t_{13} **= 40**	t_{14} **= 10**	t_{15} **= 80**	t_{15} **= 30**	t_{16} **= 10**	***r* = 1**	
	10^3	106.6 (6.6%)	74.7 (6.7%)	21.3 (6.5%)	42.6 (6.5%)	10.7 (7.0%)	85.4 (6.8%)	32.0 (6.7%)	10.7 (7.0%)	0.96 (–4.0%)	0
	10^4	106.5 (6.5%)	74.4 (6.3%)	21.4 (7.0%)	42.6 (6.5%)	10.8 (8.0%)	85.5 (6.9%)	32.2 (7.3%)	10.9 (9.0%)	0.96 (–4.0%)	0
	10^5	103.8 (3.8%)	73.1 (4.4%)	22.3 (11.5%)	42.6 (6.5%)	12.2 (22.0%)	83.4 (4.3%)	32.5 (8.3%)	12.2 (22.0%)	1.00 (0.0%)	0
	10^6	79.5 (-20.5%)	59.4 (-15.1%)	26.0 (30.0%)	39.4 (-1.5%)	19.3 (93.0%)	66.3 (-17.1%)	33.2 (10.7%)	19.1 (91.0%)	1.54 (54.0%)	68
	Variable	78.6 (-21.4%)	54.7 (-21.9%)	29.3 (46.5%)	31.1 (-22.3%)	7.7 (-23.0%)	61.7 (-22.9%)	23.1 (-23.0%)	9.3 (-7.0%)	1.30 (30.0%)	6

Note.− Time estimates are in My and rate estimates in $\times 10^{-3}$ s/s/My. First row (in bold) in each phylogeny denotes the true node ages and rate. Variable means that *N* varies among lineages as described in Figure 4.5. P is the percentage of the gene trees that do not match the species tree averaged across all replicates.

4.4 Divergence times of four hominoid species

We now explore the discrepancies in time and rate estimates when they are estimated under the multi-species coalescent vs. estimates obtained when ignoring the coalescent in a real data set. We use the hominoid phylogeny of Figure 4.6 as a case study. The molecular data are from Burgess and Yang (2008) and consist of 14,663 neutrally evolving loci. First we estimated the divergence times and rate ignoring the coalescent process, that is, by using the program MCMCTREE. Then we re-analysed the data under the multi-species coalescent, that is, by using the program BPP.

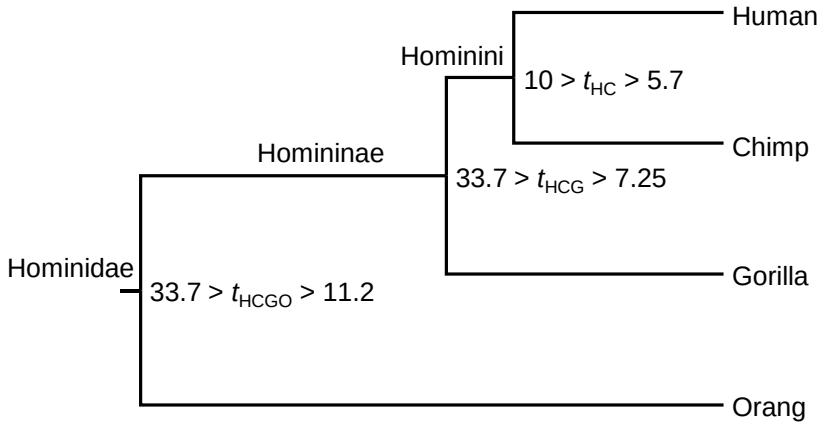

Figure 4.6: The phylogeny of four hominoid species showing the fossil calibrations used for time estimation with the program MCMCTREE. The fossil calibrations are soft, i.e., there is a 2.5% probability that the divergence time lies outside the bounds.

For the MCMCTREE analysis all loci were concatenated into a single alignment and the analysis was performed under the JC69 substitution model and the strict clock. The time unit was set to 10 My. We used a diffuse gamma prior $r \sim G(1, 100)$ with mean 0.01, meaning 10^{-9} s/s/y. The fossil calibrations are from dos Reis et al. (2012) and are shown in Figure 4.6. We used an upper bound of 33.7 Ma for the human-gorilla split while only a minimum bound had been used by dos Reis et al. (2012).

In the BPP analysis the multi-locus sequence data were analyzed under the multispecies coalescent model assuming the same mutation rate across loci. A gamma prior was used for the population size parameters, $\theta \sim G(2, 500)$, with mean 0.004. The relative age of the root in the species tree (τ_{HCGO}) was assigned a gamma prior $G(4, 219)$, with mean 0.018, while the

other relative divergence times were assigned a diffuse Dirichlet prior conditioned on τ_{HCGO}. The prior mean for τ_{HCGO} was set based on a divergence time for the human-orangutan split of 18.3 Ma (Steiper and Young 2006) and a mutation rate equal to 10^{-9} s/s/y. The relative divergence times obtained with BPP were translated into geological times. We used two calibration strategies to do so: (1) Values for the age of the human-chimp divergence, t_{HC}, were sampled from a uniform distribution between 5.7 and 10 Ma (equal to the fossil calibration for this node used with MCMCTREE). Then the sampled values were used to calculate samples for the ages of the other nodes and the molecular rate (e.g. $r = \tau_{HC}/t_{HC}$ and $t_{HCG} = t_{HC} \times \tau_{HCG}/\tau_{HC}$). (2) Alternatively, values for the molecular rate, r, were sampled from a gamma distribution $G(100, 200)$ with mean 0.5 and 95% prior interval 0.4-0.6 (meaning 0.4 to 0.6 × 10^{-9} s/s/y). This distribution is based on experimental estimates of *de novo* mutation rates in the human genome (see Scally and Durbin 2012). The sampled rate values were then used to calculate the divergence times (i.e. $t_{HC} = \tau_{HC}/r$). Similarly, we obtained estimates of ancestral population size (N) by sampling from θ and assuming a generation time of 20 years for the ancestral hominoid lineages (Langergraber, et al. 2012; Scally and Durbin 2012).

The MCMC algorithm in the MCMCTREE program was run for burn-in 10^6 iterations and collected 10^4 samples from the posterior, sampling every 2,000 iterations. In the BPP a burn-in of 10^5 iterations was used and we collected 10^4 samples from the posterior sampling every 100 iterations. The MCMC algorithm in both programs was run twice with different starting values to assess convergence.

Bayesian estimates of divergence times and the molecular rate obtained with the MCMCTREE and BPP programs are shown in Table 4.4. The posterior mean of the molecular rate obtained with MCMCTREE , $\tilde{r}$ = 0.80 × 10^{-9} s/s/y, is higher than the estimate obtained with BPP, $\tilde{r}$ = 0.53 × 10^{-9} s/s/y, when the human-chimp split is used to calibrate the phylogeny (Table 4.4). The BPP estimate is well within the 0.4 × 10^{-9} to 0.6 × 10^{-9} s/s/y range from mutation experiments (Scally and Durbin 2012).

Note that the uncertainties (or relative errors) of the MCMCTREE calibrations are 54.8%, 129% and 100% for the human-chimp, the human-gorilla, and the human-orangutan calibrations, respectively, where the uncertainty is measured as the (calibration width)/(calibration midpoint) (dos Reis and Yang 2013a). Thus, the human-chimp calibration is by far the most precise, and it is thus the most informative to estimate the molecular rate. Given that there is substantial ancestral polymorphism and incomplete lineage sorting in the ape phylogeny (Burgess and Yang 2008), the estimated molecular rate by MCMCTREE is almost surely an overestimate. We can use the estimate of equation (4.14) to gain insight into the overestimation error on the molecular rate. For example,

Table 4.4: Posterior means and 95% CIs of divergence times, rate and population sizes for the hominoid phylogeny.

	MCMCTREE	**BPP**	**BPP**		**BPP**
		t_{HC} ~ U(5.7, 10)	r ~ G(100, 200)		No calibration
t_{HCGO}	22.9 (16.3, 28.0)	26.6 (19.5, 33.4)	27.8 (22.6, 33.5)	τ_{HCGO}	13.7 (13.6, 13.9)
t_{HCG}	10.9 (7.8, 13.4)	12.8 (9.5, 16.2)	13.4 (10.9, 16.2)	τ_{HCG}	6.6 (6.6, 6.7)
t_{HC}	8.3 (5.9, 10.1)	7.9 (5.8, 9.9)	8.2 (6.6, 9.9)	τ_{HC}	4.1 (4.0, 4.2)
r	0.80 (0.63, 1.08)	0.53 (0.41, 0.69)	0.50 (0.40, 0.60)	-	-
N_{HCGO}	-	197 (144, 247)	205 (165, 247)	θ_{HCGO}	8.1 (7.8, 8.4)
N_{HCG}	-	85 (62, 106)	88 (71, 106)	θ_{HCG}	3.5 (3.4, 3.6)
N_{HC}	-	147 (106, 190)	154 (123, 186)	θ_{HC}	6.1 (5.7, 6.5)

Note.− Estimates are the posterior means and 95% CIs (in brackets). Divergence times are in My. The rate r is in 10^{-3} s/s/My. The θ (= $4Nrg$) and τ (= rt) parameters are scaled by 10^3. The population sizes, N, are in 10^3. To calculate N, a generation time of 20 years was assumed.

assuming a true mutation rate of 0.5×10^{-9} s/s/y (Scally and Durbin 2012), a true divergence time for human-chimp of 7.85 Ma (the midpoint of the calibration), an ancestral population size of 150,000 individuals (Table 4.4), and a generation time of 20 y (Langergraber, et al. 2012), we get that the rate estimate (equation 4.14) ignoring the coalescent process is $\hat{r}' \approx 0.88 \times 10^{-9}$ s/s/y, which is reasonably close to the posterior mean of $\tilde{r} = 0.8 \times 10^{-9}$ s/s/y obtained with MCMCTREE.

Time estimates for the human-gorilla and human-orangutan divergences obtained with MCMCTREE are substantially younger than those obtained with BPP. For example, the posterior mean of the human-orangutan divergence obtained with MCMCTREE at 22.9 Ma is 14% and 18% younger than those obtained with BPP at 26.6 Ma and 27.8 Ma (Table 4.4). Thus the MCMCTREE estimate of the root age is likely an underestimate. Note that time estimates obtained with BPP under the rate calibration are the most precise (i.e. they have the narrower 95% CI width). This is because the rate calibration has less uncertainty than the human-chimp calibration: 40% vs. 54.8% respectively. Assuming that the rate calibration is correct (i.e. that the mutation rate measurements are accurate) then the time estimates under the BPP rate calibration would be the most accurate and should be preferred.

4.5 Remarks and conclusions

Results from the theoretical, simulation, and real data analyses indicate that polymorphism in ancestral lineages and incomplete lineage sorting can significantly affect

Bayesian estimates of divergence times and of the molecular evolutionary rate when the inference models do not account for the multi-species coalescent. This is the case for both shallow and old phylogenies with the biases to be higher in recent divergences. The biases in time and rate estimates are more significant in case of large population sizes relative to the species divergence times where conflicts among gene trees are favoured. Whether times are over or underestimated depends on the relative precision and configuration of the fossil calibrations on the tree. If very precise calibrations are used on young nodes on a phylogeny, the ages of ancient divergence times can be grossly underestimated. Note that this is expected to occur even in ancient phylogenies. For example, if in a three species phylogeny the age of the young node, t_2, and the ancestral population size are such that the gene divergence time is twice the age of the young node ($2Ng = t_2$), then the molecular evolutionary rate will be overestimated by 100% (i.e. it will be roughly twice the true value), when the age of the young node is used as the calibration time. The age of the root (t_1) will be underestimated by approximately 50% (i.e. it will be approximately half the true value) if the true root age is much older than the age of the young node ($t_1 \gg t_2$). On the other hand, if the most precise calibrations are placed on the most ancient nodes of a phylogeny (perhaps a less common case), then the ages of the younger nodes in the phylogeny will tend to be overestimated. In both cases the molecular rate will tend to be overestimated.

Note that in our Bayesian analyses with the MCMCTREE program the sequence data were analysed as a single concatenated alignment. Alternatively we could separate each locus into individual partitions (or group into several partitions) and estimate the divergence times assuming variable rates among loci. This approach is not expected to affect time estimates and their errors because inference is done under the strict molecular clock and the species phylogeny is assumed known and the same for all loci. Indeed when we re-analysed the simulated data in the nine-species phylogeny with MCMCTREE by allowing each locus to evolve according to its own substitution rate the time estimates were virtually identical to those for the concatenated alignment. On the other hand, in the new analyses we obtained individual rate estimates for each locus, with the rate estimates being overestimated and following a distribution centred around the single rate estimate for the concatenated analysis.

Here we assumed that species were completely separated after speciation, with no gene flow between the novel species after the speciation event and with incomplete lineage sorting to be the only aspect of evolution causing incongruencies among the gene trees. This is clearly an unrealistic assumption and the effect of this on divergence time estimates requires further work (e.g. Leache, et al. 2014). An additional assumption of the multi-species coalescent is that the sequences sampled are neutrally evolving, like the set of sequences analysed for the hominoid phylogeny (Burgess and Yang 2008). Episodes of positive

selection may affect the relative ages of gene coalescent events and may affect divergence time estimates. More work will be required to address this issue.

Although incomplete lineage sorting and ancestral polymorphism have long been regarded as important aspects of molecular evolution (Takahata, et al. 1995; Edwards and Beerli 2000; Knowles and Kubatko 2010; Yang 2014), the vast majority of molecular clock dating studies have ignored this issue (e.g. dos Reis, et al. 2012; Jarvis, et al. 2014), perhaps under the belief that incomplete lineage sorting and ancestral polymorphism should only be taken into account when analysing closely related species. The results presented here highlight that the problem is much worse and that the coalescent process should be incorporated into analyses of divergence times at all timescales. Unfortunately, software currently available to perform Bayesian phylogenetic inference under the multi-species coalescent is either computationally expensive (e.g. *BEAST, Heled and Drummond, 2010) or has been designed to work only for closely related sequences (e.g. BPP, Yang, 2015), thus restricting such analyses for small phylogenies with only a few taxa. For example, we chose to analyze the hominoid phylogeny because the BPP program can only perform inference under the strict molecular clock and under the JC69 substitution model. These assumptions are met in the hominoid phylogeny: The clock is not violated and the molecular distances are small enough so that the JC69 model can adequately describe the substitution process. In order to analyze more ancient phylogenies, multi-species coalescent models that incorporate molecular rate variation among lineages, that use more complex substitution models, and that can handle the large amounts of genomic data now available will be required.

5 An evaluation of different partitioning strategies for Bayesian estimation of species divergence times

In the previous chapter we explored the performance of two Bayesian algorithms in estimating species divergence times from molecular data when ancestral polymorphism and incomplete lineage sorting are present in the data. In this chapter we will study the effect of five commonly used partitioning strategies in Bayesian estimation of species divergence times by analysing simulated and real data sets.

5.1 Accounting for heterogeneity in evolutionary substitution patterns

It is well recognised that different parts of the genome may evolve at different rates and with different patterns of substitution (Springer, et al. 1999; Shapiro, et al. 2006). With large molecular data sets typically analyzed in phylogenetic studies (Meusemann, et al. 2010; dos Reis, et al. 2012; Jarvis, et al. 2014; Misof, et al. 2014; Ruhfel, et al. 2014) there is an increased need to adequately model the underlying patterns of evolution (Brown and Lemmon 2007). The gamma rate-heterogeneity model is such an approach as it relaxes the assumption of rate constancy across sites (Yang 1994c). Although such models offer a significant improvement in phylogenetic inference (Yang 1996a; Sullivan and Swofford 2001), they assume the same substitution patterns for all sites and might not perform well. For example, the first, second and third codon positions might have different substitution patterns owing to different selection pressures, independently of having the same evolutionary rate or not.

Several methods have been proposed to model the heterogeneity in rates and patterns of substitution across the sites in a sequence alignment. These include mixture models (Koshi and Goldstein 1995; Pagel and Meade 2004; Le, et al. 2008; Lartillot, et al. 2009) and partitioning (Yang 1996b; Koshi and Goldstein 1998; Nylander, et al. 2004; Brandley, et al. 2005; Brown and Lemmon 2007). In mixture models each site is assigned to a substitution model with a probability, with the model parameters and their probabilities to be estimated from the data. In partitioning, the alignment is first divided into several site partitions and then independent substitution models are assigned for each partition.

The major problem in a partition analysis is to choose an appropriate partitioning scheme; that is to divide the alignment into partitions consisting of sites with similar evolutionary histories. One suggested approach is to select a partitioning scheme for a given data set according to a statistical criterion such as the Bayesian Information Criterion (BIC). However, even for very short alignments the number of possible partitioning schemes is too large to be computationally possible to evaluate all of them using the BIC (Li, et al. 2008; Lanfear, et al. 2012). A typical approach for researchers is to specify a partitioning scheme based on the structural features of the sequences in the alignment. This often results in defining partitions on the basis of genes, codon positions, coding, non-coding, mitochondrial or nuclear regions (Strugnell, et al. 2005; Shapiro, et al. 2006; dos Reis, et al. 2012; Fong, et al. 2012; Jarvis, et al. 2014). Recently, algorithmic approaches have been developed which start from user-defined sets of sites (data blocks) and iteratively merge the sets that improve the most the score of an information criterion at each step, until there is no further improvement in the score (Lanfear, et al. 2012; Lanfear, et al. 2014). This approach uses heuristic algorithms to reduce the total number of schemes to evaluate and return the best-fitting scheme among a subset of all possible partitioning schemes. Such an algorithm has been implemented in the PartitionFinder program and is computationally efficient even for very large data sets (Lanfear, et al. 2014). The underlying assumption, however, is that all the sites within the user-specified data blocks have evolved similarly. A recent change, introduced by Frandsen et al. (2015), is to treat all sites of the alignment as a single focal subset and use an iterative *k*-means algorithm to divide the focal subset into finer partitions according to site rates, making the assignment of data blocks unnecessary.

The choice of partitioning scheme may affect any downstream phylogenetic analysis such as phylogeny reconstruction or estimation of species divergence times and rates. Several studies have examined the effect of partitioning scheme on the inference of topology (Strugnell, et al. 2005; Ward, et al. 2010; Xi, et al. 2012; Leavitt, et al. 2013) and it has been observed that underpartitioning might lead to significant bias, as for example highly supported but incorrect nodes on the estimated tree (Kainer and Lanfear 2015). However, there has been no systematic effort to explore the effect of partitioning on species divergence time estimation. Two studies from Poux et al. (2008) and Voloch and Schrago (2012) used only closely related species with many calibrations and they found practically no differences among partitioning schemes. Zhu et al. (2015) suggest that in relaxed clock dating, increasing the number of partitions is very important to improving the precision of divergence time estimation when the fossil calibration information is fixed. An important aspect in dating studies is the pattern of rate variation among the branches of the tree topology, which may vary across the alignment. The use of some partitioning schemes may fail to accommodate such variation, which may result into poor time estimates. Thus, when

estimating species divergence times there is a need for a proper data partitioning scheme which will capture the variation of substitution patterns, absolute rate and among-branches rate variation across the alignment.

Below we explore the performance of five commonly used partitioning schemes on Bayesian estimation of species divergence times using simulated and real data. We simulate sequence alignments from a nine-species phylogeny with known node ages and analyse them to estimate the divergence times using the five partitioning schemes. We study two different cases of clock violation (slight and severe violation of the clock) and use various prior assumptions. Results indicate that the time estimates are very similar among schemes, especially when the clock is not seriously violated. Highly partitioned schemes reduce uncertainty of posterior time estimates but may lead to biases under incorrect prior assumptions. However, some results are unexpected and in the absence of any clear explanation any further safe conclusion is precluded.

5.2 Methods

5.2.1 Design of the simulation experiment

We used the nine-species phylogeny of Figure 5.1 to simulate 50 gene alignments and examine how the species divergence time estimates vary across five commonly used partitioning strategies. Each gene alignment is simulated using the species tree, with the ages of the nodes to be $t_1 = 1$, $t_2 = 0.95$, $t_3 = 0.55$, $t_4 = 0.40$, $t_5 = 0.25$, $t_6 = 0.15$, $t_7 = 0.10$, $t_8 = 0.50$ (Figure 5.1). The time unit is set to 100 million years (My) so, for example, the ages of nodes 1 and 2 are 100 Ma and 95 Ma, respectively. We assume a mean substitution rate $\mu_0 = 0.5$ (meaning 0.5 substitutions per site per time unit or 5×10^{-9} substitutions per site per year) over all genes and lineages. We set the overall rate across lineages for gene g to be a random variable from the gamma distribution $\mu_g \sim G(10, \frac{10}{\mu_0})$, with mean μ_0 and 95% interval (0.24, 0.85). The log-rates for the branches of the gth gene tree are generated as independent random variables from the normal distribution $\log\mu_{gb} \sim N(\log\mu_g - \sigma^2/2, \sigma^2)$, for $b = 1,...,16$. Thus, the branch rates for the gth gene are independent random variables from a lognormal distribution with mean μ_g. Then for the gth gene tree we multiply the μ_{gb} with the time duration of the bth branch to calculate the branch length. In this way we construct 50 gene trees with branch lengths. We use two values for the variance $\sigma^2 = 0.01$ and 0.25,

corresponding to slight and serious violation of the clock, respectively. In either case the 50 genes may have different overall rates, but all genes have the same extent of among-branches rate variation (the same σ^2). Simple R code is written to sample the branch rates and generate the gene trees.

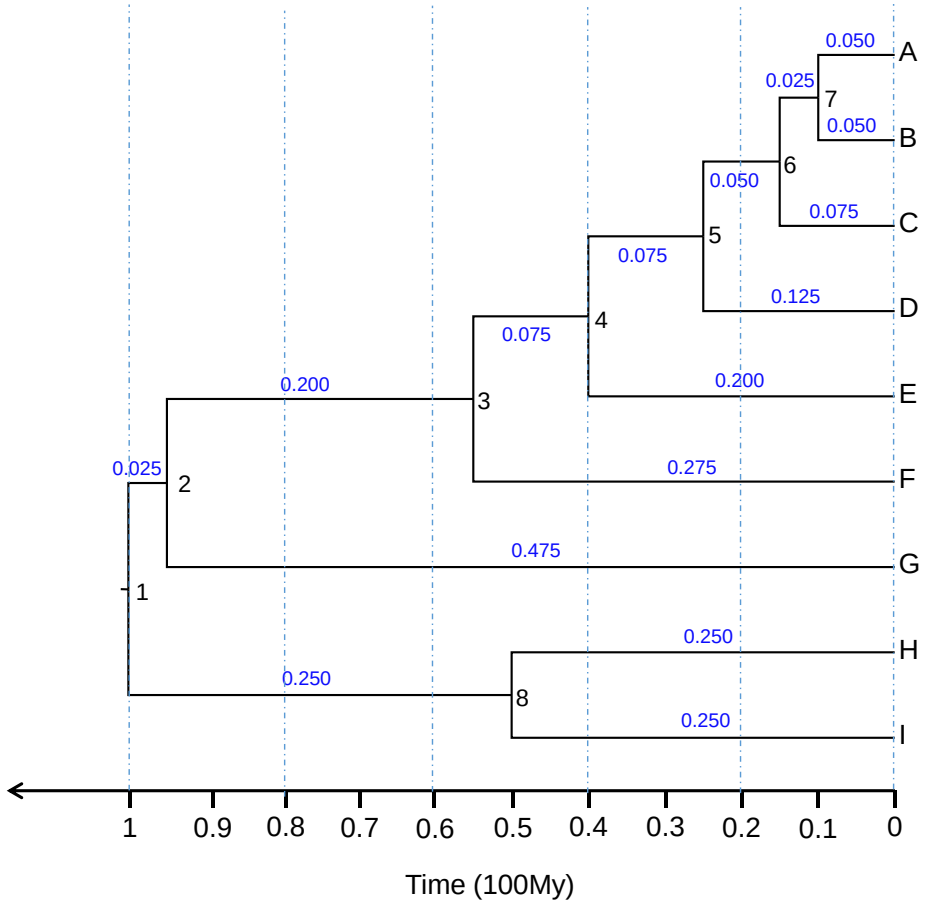

Figure 5.1: Species tree used to simulate gene alignments. Internal nodes are numbered from 1 to 8. Branch lengths, assuming a substitution rate 5×10^{-9} substitutions per site per year for all branches, are shown in blue.

The generated gene trees have branch lengths measured in substitutions per site. Because protein-coding genes are widely used in dating studies (Meusemann, et al. 2010; dos Reis, et al. 2012; Misof, et al. 2014) we simulate gene alignments using a codon model. Thus we multiply all branch lengths in all gene trees by 3. Gene alignments are then simulated on the gene trees under the M3 (discrete) model of codon evolution (Yang, et al. 2000a) using the program EVOLVERNSSITES from PAML4.8 (Yang 2007). This model allows for three classes of codons with different nonsynonymous to synonymous rate ratio $\omega_0 = 0.01$, $\omega_1 = 0.5$ and $\omega_2 = 0.9$. We simulate 25 conserved genes with probabilities $p_0 =$

0.8, $p_1 = 0.19$, $p_2 = 0.01$ for the three site classes, with the average ω to be 0.112, and 25 non-conserved genes with probabilities $p_0 = 0.5$, $p_1 = 0.3$, $p_2 = 0.2$, with the average ω to be 0.335. The sequence length of each gene is $n = 500$ codons, the transition/transversion rate ratio is $\kappa = 2$ and the codon frequencies are assumed to be equal. The number of replicates is 100. Thus we simulate 2 × 100 data sets, each consisting of 50 genes, with 100 for $\sigma^2 = 0.01$ and 100 for $\sigma^2 = 0.25$.

5.2.2 Estimation of divergence times from the simulated gene alignments

We analyzed the simulated gene alignments with the program MCMCTREE v4.8 (Yang 2007) to estimate the species divergence times. We used the following five partitioning schemes:

1) We concatenated all genes into a single "supergene" (C).
2) We concatenated the 1st and 2nd codon positions from all genes into one partition and the 3rd codon positions from all genes into another (CP).
3) We used the program PartitionFinder v1.1.1 (Lanfear, et al. 2012; Lanfear, et al. 2014), (PF), with codon positions 1+2 and 3 of each gene treated as a data block. The program explores different partition strategies using the BIC. The number of possible partitions ranges from 1 to 100.
4) We analyzed the data as 50 partitions with each partition to be a gene alignment (G).
5) We treat the 1st and 2nd codon position of each gene as a partition and the 3rd codon positions as another, creating in total 2 × 50 = 100 partitions (GCP).

In the PartitionFinder program the user has to provide data blocks. Different data blocks may be merged (concatenated) into one partition, but sites in the same data block will never be separated into different partitions. The program estimates the best-fitting partitioning scheme and the best-fitting substitution model for each partition from a user-specified set of models based on an information criterion. The topology is either provided by the user or estimated by the data.

We used the data blocks defined in the GCP scheme as the starting point and the tree of Figure 5.1. We did not search for the best-fitting substitution model for each partition but we used the HKY85 + Γ substitution model for each partitioning scheme (Hasegawa, et al. 1985). Automatic model selection techniques (Posada and Crandall 1998) might suggest the use of pathological models, such as the 'I + Γ' models, or models unavailable to other programs, or parameter-rich models which might lead to over-fitting if applied to small

partitions. Furthermore, MCMCTREE does not allow for different substitution models across partitions and apart from that, the use of different substitution models for the same data blocks complicates the comparison of different partitioning strategies. We used the *greedy* heuristic algorithm with the BIC score to search for the best scheme since it was found to perform better than the others (*rcluster*, *hcluster*), although it requires more computations (Lanfear, et al. 2014). We also used the *linked* option for the branch length estimation according to which one set of branch lengths is estimated and a scaling parameter is used to adjust the branch lengths within each partition.

We set the time unit in MCMCTREE to be 100 My and applied three calibration strategies: 1) We assigned the prior $t_1 \sim B(0.8, 1.2)$ to the root age, meaning a uniform distribution between 0.8 and 1.2 with left and right tail probabilities 2.5% that the age of the root is outside the bounds (Yang and Rannala 2006). This mimics a soft-bound calibration at the root age between 80 Ma and 120 Ma based on the fossil record. 2) We applied the same constraint in the root age and the constraint $t_3 \sim B(0.525, 0.575)$ to the age of node 3. This mimics a weak calibration on the root and an informative calibration in the younger node 3. 3) The same constraint in the root age and a conflicting constraint $t_3 \sim B(0.575, 0.625)$ on the age of node 3. Note that the true age of node 3 is outside those bounds. This mimics an incorrect calibration on node 3. The prior ages of the uncalibrated internal nodes are specified from a birth-death process through a uniform kernel with rates $\lambda = \mu = 1$ and $\rho = 0$ (Yang and Rannala 2006). We ran the MCMCTREE program without data for calibration strategies 2 and 3 and found that the marginal time priors in the calibrated nodes closely matched the user-specified densities.

The rates at the different partitions (loci) are assigned the gamma-Dirichlet prior (dos Reis, Donoghue, et al. 2014). A gamma prior is assigned to the average rate over all loci ($\bar{\mu}$) and the overall locus rates are then calculated based on the uniform Dirichlet distributions with parameter $\alpha = 1$. We used $\bar{\mu} \sim G(2, 4)$ with mean 0.5, meaning 5×10^{-9} substitutions per site per year with prior 95% interval (0.6, 14.0). The mean of this prior matches the overall substitution rate ($\mu_0 = 0.5$) of all genes under which the gene alignments were simulated. We also used two "incorrect" priors, to assess the performance of the partitioning schemes under incorrect rate priors: (i) a slow rate, $\bar{\mu} \sim G(2, 40)$ and (ii) a fast rate $\bar{\mu} \sim G(2, 0.4)$. To model the among-branches rate variation we used the independent-rates model (IR) which matches the way the data were simulated. We also used the autocorrelated-rates model (AR) to mimic a scenario where the rate-drift model used does not describe properly the among branches rate heterogeneity of a data set. A gamma prior, $\bar{\sigma}^2 \sim G(2, 20)$, was assigned to the average rate drift parameter among loci with the locus-specific parameters to be defined from the Dirichlet process. The topology of Figure 5.1 was used along with the HKY85 + Γ_4 model of

nucleotide substitution. The approximate likelihood method was used for computational efficiency (dos Reis and Yang 2011). The same settings were used in all replicates.

The MCMC was run with a burn-in of 10^6 steps and collecting 10^4 samples from the posterior every 500 steps. For the partitioning strategy GCP posterior samples were collected every 250 steps to save computational time. Convergence was evaluated for only the first replicate for each combination of rate prior, calibration strategy, rate-drift model and partitioning scheme by running two independent MCMC runs with different starting values. For each replicate we estimated the posterior time means and the 95% HPD intervals. Those are averaged over the 100 replicates.

5.2.3 Evaluating the performance of partitioning strategies

We use the following measures to evaluate the performance of partitioning strategies in terms of accuracy and precision of the posterior time estimates. For each partitioning scheme we average the first three measures across all nodes and replicates.

(i) Average Relative Error. We calculate the relative error of the time estimate in node *i* in the *j*th replicate $d_{ij} = \left| \frac{\tilde{t}_{ij} - t_i}{t_i} \right|$, where t_i is the true age of the node *i* and $\tilde{t}_{ij}$ is the time estimate (posterior mean), with $j = 1,...,100$, $i = 1,...,s-1$, where *s* is the number of species. This may be considered a measure of accuracy.

(ii) Relative HPD Width. We calculate the relative HPD interval width of the time estimate for node *i* in the *j*th replicate as $sw_{ij} = \frac{w_{ij}}{t_i}$, where w_{ij} is the 95% HPD interval of the time estimate $\tilde{t}_{ij}$. This is a measure of precision.

(iii) Mean Square Error (MSE). The root of MSE of the time estimate of node *i* in the *j*th replicate is $\sqrt{\text{MSE}_{ij}} = \sqrt{Var(\tilde{t}_{ij}) + (\tilde{t}_{ij} - t_i)^2}$. The $Var(\tilde{t}_{ij})$ is estimated as $\left(w_{ij} / (2*1.96) \right)^2$. This is a measure of both accuracy and precision of the time estimates.

(iv) Coverage Probability. For each partitioning scheme we calculate the percentage that the HPD interval of the time estimate for node *i* contains the true age t_i, from the *R* replicates (P_i^k). Then we average over all the nodes $\bar{P} = \frac{1}{s-1} \sum_{i=1}^{s-1} P_i$.

5.2.4 Plants data set

We estimated the divergence times of fifteen plant species using the five partitioning schemes considered in the 2 simulations. The molecular data are from Ruhfel et al. (2014) and consist of 78 plastid gene alignments (58,347 sites in total). The phylogeny with fossil calibrations is shown in Figure 5.2.

We used PartitionFinder1.1.1 with the same settings as in the simulation analysis, except that the GTR + Γ nucleotide substitution model was used. Note that here the G scheme involves 78 partitions (one for each gene) while GCP involves 156 (one for each gene and codon position). The program MCMCTREE4.8 was used for the Bayesian dating analysis with one time unit to be 100 My. We used three priors for the average rate, $\bar{\mu} \sim G(1, 100)$, $\bar{\mu} \sim G(1, 10)$ and $\bar{\mu} \sim G(1, 1)$ (Magallon, et al. 2013b). The first prior specifies slow rates with mean rate 10^{-10} substitutions per site per year while the third specifies fast rates with mean rate 10^{-8}. The time priors were constructed from the calibrations of Figure 5.2 together with the birth-death process, with rates $\lambda = \mu = 1$, and $\rho = 0$. For the rate drift parameter we used the prior $\bar{\sigma}^2 \sim G(1, 10)$. The GTR + Γ_4 substitution model was used in all partitions and approximate likelihood calculation was used to save computational time. We used both the independent and the autocorrelated rates model for the among-branches rate variation.

All MCMC analyses were run with the same settings as in the simulation. For each combination of rate prior, rate-drift model and partition scheme the MCMC was run twice from different starting values to evaluate convergence.

5.3 Results

For the simulated data we estimated the species divergence times using five different partitioning schemes: 1) concatenation (C), 1 single partition; 2) codon positions (CP), 2 partitions (codon positions 1+2 vs. 3); 3) PartitionFinder (PF); 4) gene (G), 50 partitions; and 5) both gene and codon positions (GCP), 100 partitions. In case of clock-like genes the number of partitions determined by PartitionFinder varied from 9 to 16 among replicates, while for the non clock-like genes it was from 9 to 17. We evaluated the performance of the partitioning schemes using four performance measures: relative error, relative HPD width, mean square error and coverage probability, under three calibration strategies, three rate priors and two models of among-branches rate variation.

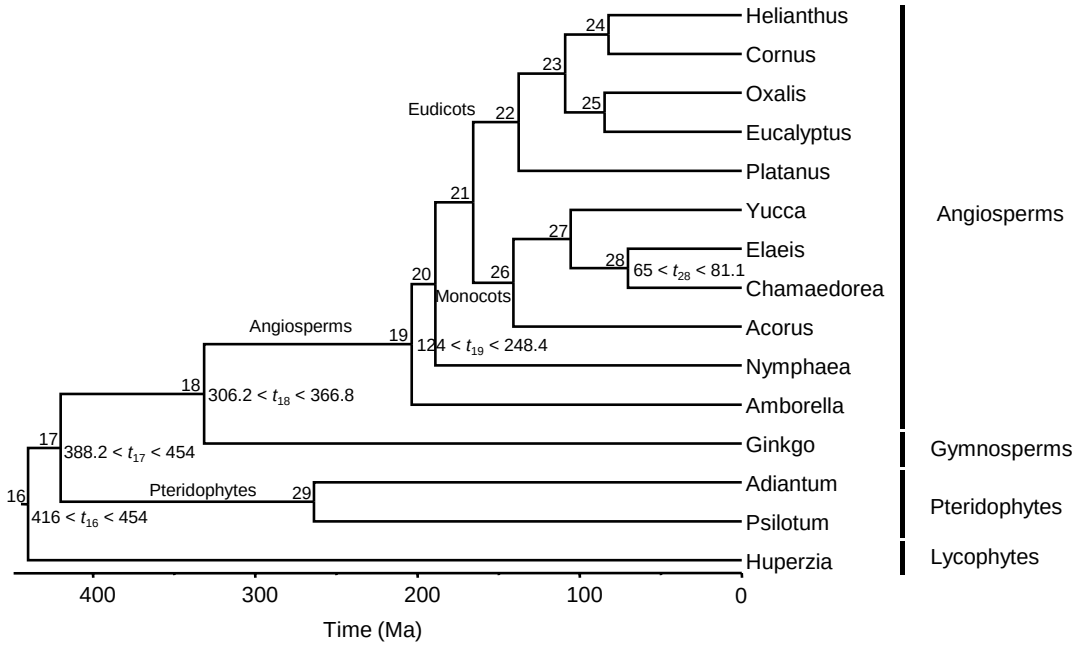

Figure 5.2: Phylogeny of 15 plant species. The relative positions of the major groups (lycophytes, pteridophytes, gymnosperms and angiosperms) are according to Magallon et al. (2013b) while those within the angiosperms are from Ruhfel et al. (2014). Nodes are numbered from 16 to 29. Fossil calibrations for five nodes are shown next to the nodes. The fossil bounds are soft, with 5% probability that the true age is outside the bounds (2.5% probability in each site). The calibrations are according to Clarke et al. (2011) and Zanne et al. (2014).

5.3.1 Results from simulations when the clock is seriously violated

In data sets simulated using $\sigma^2 = 0.25$ the molecular clock is seriously violated. When there is a single calibration t_1~B(0.8, 1.2), the rate prior is $\bar{\mu} \sim G(2, 4)$, and the independent-rates model is used, the time estimates are close to their true values for all partitioning schemes (Figure 5.3B). The relative errors averaged over all nodes and replicates are 0.028, 0.046, 0.048, 0.039 and 0.039 for the partitioning schemes C, CP, PF, G, and GCP, respectively (Table 5.1). The differences in time estimates among the partitioning schemes are small; however, the C scheme seems to be the most accurate. The age of the root is estimated more accurately than the age of all other nodes for all partitioning schemes due to the single calibration on the root. The precision of time estimates under the C scheme is lower than that of all other schemes and the true ages are well within the HPD time intervals

for all partitioning schemes (Table 5.1). However, the G and GCP schemes seem to perform best with respect to MSE ($\sqrt{\text{MSE}}$ = 0.056; Table 5.1).

When we add another good fossil calibration on the internal node 3, $0.525 < t_3 < 0.575$ for the true age $t_3 = 0.55$, the time estimates become more precise for all nodes and partitioning schemes (compare Figure 5.3B' with 5.3B). For example the relative HPD width over all nodes for the C scheme reduces from 0.50 to 0.26 and from 0.43 to 0.16 for the G scheme (Table 5.1). Accuracy is either the same or improved for the partitioning schemes C, CP and PF but is slightly worse for the highly partitioned schemes G and GCP (Table 5.1). The age of node 3 is accurately and precisely estimated for all partitioning schemes owing to the informative calibration on it, whereas the age of the root is not accurately estimated in all partitioning schemes. For example, the relative error for the root is increased from 0.002 to 0.029 for the C scheme and from 0.003 to 0.030 for the G scheme after the inclusion of the additional calibration in node 3 (Table 5.2). All partitioning schemes but the C scheme have similar MSE (e.g. $\sqrt{MSE}$ = 0.026, 0.026, 0.027, 0.025 for the schemes CP, PF, G, GCP, respectively) but the highly partitioned schemes G and GCP have smaller coverage probabilities due to the higher precision and lower accuracy that they have.

We now explore the impact of partitioning schemes when incorrect rate priors are used; a slow rate $\bar{\mu} \sim G(2, 40)$ or a fast rate $\bar{\mu} \sim G(2, 0.4)$. When using a single calibration on the root the time estimates are worse using the slow rate prior than using the correct rate prior for all partitioning schemes (Figure 5.3A, B). The time estimates under the slow rate prior are older and more biased than those under the correct prior. For example, the absolute relative errors with the slow prior are 0.177 and 0.203 with the C and CP schemes compared to 0.028 and 0.046 with the correct rate prior (Table 5.1). Moreover, the use of the slow rate prior produces misleadingly precise estimates (Tables 5.1 and 5.2), since the estimates are far from the true values and for many nodes the true ages are not within the HPD intervals for all partitioning schemes (Figure 5.3A). For example, the age of the root was estimated at 114.6, 115.7, 116.0, 115.0 and 116.4 for the schemes C, CP, PF, G and GCP, respectively, with only the HPD interval using the C scheme to include the true value (in 48 out of 100 replicates; Table 5.1). Overall, all partitioning schemes seem to perform equally bad in case of a slow rate prior when a single calibration is used on the root, with the G scheme to be preferable in terms of MSE (Table 5.1). The use of another correct calibration on node 3 improves the time estimates with the slow rate prior for all partitioning schemes. The fast rate prior gives similar estimates to the correct rate prior especially when two calibrations are used (Tables 5.1 and 5.2). Hence, results for this case are not reported in Figure 5.3.

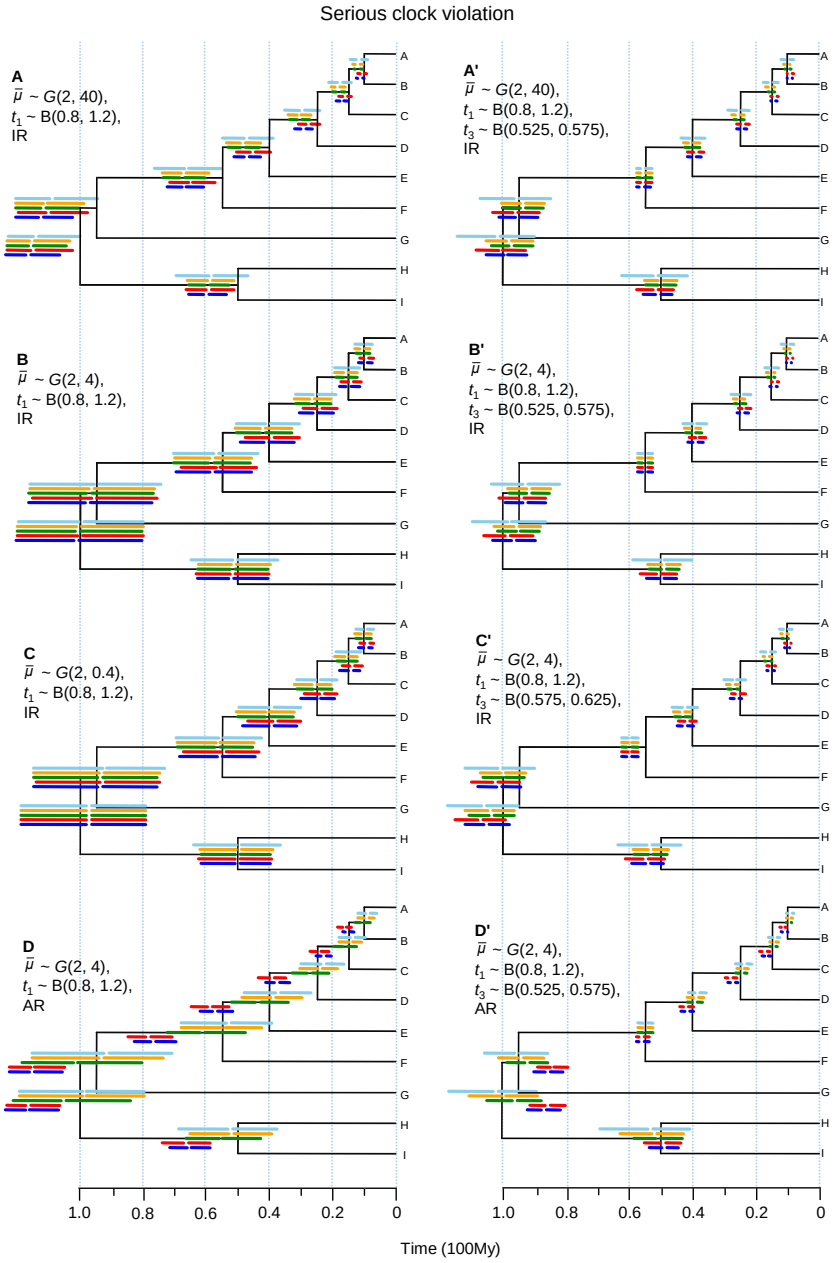

Figure 5.3: Posterior divergence time estimates from simulated data when the clock is seriously violated for each combination of rate prior, calibration strategy and rate-drift model. The true timetree is shown in black. Horizontal bars represent the 95% high posterior density intervals under the five partitioning strategies. These are (from the top to the bottom): (i) concatenation (C), 1 single partition (light blue); (ii) codon positions (CP), 2 partitions (codon positions 1+2, 3) (yellow); (iii) PartitionFinder (PF), variable partitions (green); (iv) gene (G), 50 partitions (red); and (v) both gene and codon positions (GCP), 100 partitions (blue). The gaps within the bars represent the posterior means. The time estimates and their intervals are averages over 100 replicates. IR: independent-rates model, AR: autocorrelated-rates model. The time estimates with the fast rate prior $\bar{\mu} \sim G(2, 0.4)$ and two calibrated nodes are very similar to those with the correct rate prior $\bar{\mu} \sim G(2, 4)$ and two calibrations, and are not reported here.

We now explore the time estimates in case an incorrect calibration $0.575 < t_3 < 0.625$ is used on node 3 for the true age $t_3 = 0.55$, in addition to the correct calibration on the root. The accuracy of time estimates is worse for all partitioning schemes than that when correct calibrations are used. For example, the relative error for schemes C and CP are 0.072 and 0.080, respectively, compared to 0.028 and 0.046 when a single calibration is used on the root, with the GCP scheme to have the smallest relative error (Table 5.1). The precision of time estimates is higher than that under a single calibration on the root for all partitioning schemes and similar to that under two correct calibrations with the PF and GCP schemes to have the highest precision (Table 5.1). In general, all node ages are overestimated for all partitioning schemes owing to the incorrect informative calibration on node 3. The age estimate of node 3 is most significantly affected with the HPD intervals for all partitioning schemes not to include the true node age. Overall, the GCP scheme has the highest accuracy and precision but the coverage probability is low (0.62; Table 5.1).

We also analyzed the simulated data sets with the autocorrelated-rates model (Figure 5.3D and 5.3D'). In this case, the time estimates show considerable variation among the partitioning schemes. When using a single calibration on the root, increasing the number of partitions produces older and biased time estimates for all nodes (Figure 5.3D). For example, the absolute relative error for the C scheme is 0.060 while is approximately 7 times higher (0.435) for the G scheme (Table 5.1). Moreover, the highly partitioned schemes lead to misleadingly high precision (Figure 5.3D, Table 5.1). With the addition of a correct calibration on node 3 the accuracy of time estimates is improved, particularly for the schemes G and GCP. However, the ages of the deep nodes (i.e. nodes 1 and 2) are more severely underestimated as the number of partitions increases while those of younger nodes are more severely overestimated (Figure 5.3D'). This is probably because the calibration in the internal node 3 is more informative (uncertainty 10%) than the one in the root (uncertainty 40%). No matter the calibration strategy, the node ages estimated with a highly partitioned scheme (PF, G, GCP) are seriously biased (Figure 5.3D, D') and have very small coverage probabilities (Table 5.1).

Table 5.1: Performance of different partitioning strategies in data simulated with serious clock violation.

Rate model	$\bar{\mu}$ ~	Calibration	rel. error					HPD width/age					$\sqrt{MSE}$					coverage probability (%)				
			C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)
IR	*G*(2, 40)	t_1 ~ B(0.8, 1.2)	0.177	0.203	0.204	**0.138**	0.177	0.40	0.27	0.24	0.26	**0.22**	0.092	0.096	0.098	**0.082**	0.094	**72**	7	5	46	11
	G(2, 4)	t_1 ~ B(0.8, 1.2)	**0.028**	0.046	0.048	0.039	0.039	0.50	0.45	0.44	0.43	**0.42**	0.060	0.058	0.058	**0.056**	**0.056**	100	100	100	100	100
	G(2, 0.4)	t_1 ~ B(0.8, 1.2)	**0.032**	0.036	**0.032**	0.052	0.037	0.50	0.45	0.44	**0.42**	**0.42**	0.060	0.057	**0.055**	0.057	**0.055**	100	100	100	100	100
	G(2, 40)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.525, 0.575)	**0.030**	0.032	**0.030**	0.039	0.032	0.27	0.16	**0.14**	0.16	**0.14**	0.034	0.024	0.023	0.026	**0.022**	**100**	96	95	92	93
	G(2, 4)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.525, 0.575)	**0.028**	0.030	0.032	0.046	0.040	0.26	0.16	0.15	0.16	**0.14**	0.033	0.026	0.026	0.027	**0.025**	**100**	94	91	85	86
	G(2, 0.4)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.525, 0.575)	**0.029**	0.040	0.032	0.047	0.041	0.26	0.17	0.15	0.16	**0.14**	0.033	0.029	**0.026**	0.027	**0.026**	**99**	92	90	83	84
	G(2, 4)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.575, 0.625)	0.072	0.080	0.072	0.063	**0.062**	0.26	0.17	**0.15**	0.17	**0.15**	0.047	0.040	**0.037**	0.043	0.038	**82**	52	53	68	62
AR	*G*(2, 4)	t_1 ~ B(0.8, 1.2)	0.060	**0.040**	0.091	0.435	0.388	0.53	0.47	0.45	0.30	**0.26**	0.067	**0.059**	0.069	0.161	0.150	**100**	**100**	**100**	0	0
	G(2, 4)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.525, 0.575)	0.044	0.036	**0.034**	0.101	0.083	0.25	0.19	0.16	0.15	**0.13**	0.037	0.030	**0.027**	0.051	0.045	**95**	92	92	38	43

Note.−The performance measures are averages over the 100 replicates and over the 8 internal nodes on the tree. The partitioning strategies are C: concatenation (1 partition), CP: codon position (2P), PF: PartitionFinder (V) G: gene (50P), GCP: gene and codon position (100P). IR: independent-rates model, AR: autocorrelated-rates model. Cells in bold indicate the preferable partitioning strategy according to the respective measure.

Table 5.2: Performance of different partitioning strategies to estimate the ages of nodes 1 (top) and 4 (bottom) when the clock is seriously violated.

Rate model	$\bar{\mu}$ ~	Calibration	rel. error					HPD width/age					$\sqrt{MSE}$					coverage probability				
			C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)
IR	*G*(2, 40)	t_1 ~ B(0.8, 1.2)	**0.141**	0.154	0.159	0.151	0.166	0.23	0.20	0.19	0.21	**0.17**	**0.153**	0.162	0.166	0.160	0.172	**48**	0	0	0	0
			0.183	0.216	0.218	**0.136**	0.188	0.40	0.27	0.25	0.27	**0.22**	**0.084**	0.091	0.091	0.061	0.078	**73**	4	1	56	2
	G(2, 4)	t_1 ~ B(0.8, 1.2)	**0.002**	0.005	0.010	0.003	0.015	0.40	0.40	0.40	0.40	0.40	0.102	0.102	0.102	0.102	0.102	100	100	100	100	100
			0.030	0.054	0.056	0.033	0.038	0.50	0.45	0.44	**0.43**	**0.43**	0.053	0.052	0.052	**0.046**	0.047	100	100	100	100	100
	G(2, 0.4)	t_1 ~ B(0.8, 1.2)	0.027	0.026	0.026	0.027	**0.025**	0.40	0.40	0.40	0.40	0.40	0.105	0.105	0.105	0.105	0.105	100	100	100	100	100
			0.032	0.034	**0.032**	0.051	0.033	0.49	0.45	0.44	**0.42**	**0.42**	0.053	0.049	0.047	0.049	**0.046**	100	100	100	100	100
	G(2, 40)	t_1 ~ B(0.8, 1.2)	**0.027**	0.031	0.033	0.028	0.028	0.24	0.15	**0.13**	0.15	**0.13**	0.070	0.051	0.049	0.050	**0.045**	**100**	93	89	99	96
		t_3 ~ B(0.525, 0.575)	**0.016**	0.020	0.018	0.026	0.019	0.18	0.12	0.12	0.13	**0.11**	0.020	0.016	0.015	0.018	**0.014**	**100**	99	**100**	99	**100**
	G(2, 4)	t_1 ~ B(0.8, 1.2)	**0.029**	0.046	0.049	0.030	0.039	0.23	0.15	0.14	0.16	**0.13**	0.068	0.063	0.063	**0.053**	0.055	**99**	82	68	**99**	80
		t_3 ~ B(0.525, 0.575)	0.026	**0.017**	0.019	0.040	0.029	0.19	0.13	**0.12**	0.14	**0.12**	0.022	0.016	**0.015**	0.022	0.018	**99**	**99**	**99**	90	96
	G(2, 0.4)	t_1 ~ B(0.8, 1.2)	**0.031**	0.048	0.051	**0.031**	0.041	0.23	0.15	0.14	0.16	**0.13**	0.069	0.065	0.065	**0.053**	0.057	**99**	78	61	98	79
		t_3 ~ B(0.525, 0.575)	0.027	**0.020**	**0.020**	0.042	0.031	0.19	0.14	**0.12**	0.14	**0.12**	0.023	0.017	**0.016**	0.023	0.018	**99**	**99**	**99**	84	94
	G(2, 4)	t_1 ~ B(0.8, 1.2)	0.063	0.045	**0.041**	0.073	0.054	0.22	0.15	**0.14**	0.16	**0.14**	0.087	0.063	**0.057**	0.085	0.068	**92**	85	85	57	65
		t_3 ~ B(0.575, 0.625)	0.067	0.086	0.079	**0.047**	0.062	0.20	0.14	**0.13**	0.14	**0.13**	0.034	0.037	0.035	**0.024**	0.028	**86**	22	30	78	58
AR	*G*(2, 4)	t_1 ~ B(0.8, 1.2)	0.013	**0.012**	0.055	0.166	0.168	0.40	0.40	0.37	**0.17**	**0.17**	0.103	**0.102**	0.112	0.171	0.173	**100**	**100**	**100**	0	0
			0.071	**0.041**	0.096	0.478	0.433	0.53	0.47	0.45	0.30	**0.26**	0.063	**0.052**	0.062	0.194	0.175	**100**	**100**	99	0	0
	G(2, 4)	t_1 ~ B(0.8, 1.2)	0.030	**0.029**	0.042	0.141	0.129	0.28	0.22	0.17	0.11	**0.10**	0.079	**0.065**	**0.065**	0.144	0.132	**100**	98	91	0	0
		t_3 ~ B(0.525, 0.575)	0.039	0.027	**0.020**	0.054	0.046	0.15	0.12	0.12	0.11	**0.10**	0.022	0.017	**0.015**	0.025	0.022	96	97	**99**	64	74

Note.− See caption of Table 5.1.

5.3.2 Results from simulation when the clock is slightly violated

When the clock is slightly violated the time estimates show similar trends to those under serious violation of the clock. However, for all combinations of prior specifications, the time estimates are more precise and accurate (Figure 5.4). For example, the relative error in case of a single calibration in the root with correct rate prior for the partitioning scheme C is 0.019 compared to 0.028 when the clock is seriously violated and the relative HPD width is 0.43 vs. 0.50, respectively (Tables 5.3 and 5.1). In general, the time estimates are more similar among partitioning schemes than in the case of serious clock violation.

The effect of an incorrect rate prior is the same as when the clock is seriously violated, with the slow rate prior to produce less accurate estimates than a correct one, for all partitioning schemes. When two correct calibrations are used and correct rate priors the time estimates are more precise than when a single calibration is used (Table 5.3). For example, with a single calibration in the root the relative HPD widths for nodes 1 and 4 with the C scheme are 0.40 and 0.43, while they are 0.16 and 0.13, respectively with two correct calibrations (Table 5.4). When an incorrect calibration is used in node 3 all node ages are slightly overestimated for all partitioning schemes as in the serious clock violation.

With an incorrect rate-drift model the time estimates show the same pattern as in the case of serious clock violation, although differences among partitioning schemes are smaller. With a single calibration in the root the time estimates under the partitioning schemes C, CP, PF and G are close to the true values while the time estimates under the GCP scheme are older and less accurate, especially for the deep nodes (Figure 5.4D). Adding a correct calibration on node 3 improves the time estimates for all partitioning schemes (Figure 5.4D'). However, all schemes but the C scheme tends to give younger and less accurate estimates for the deep nodes.

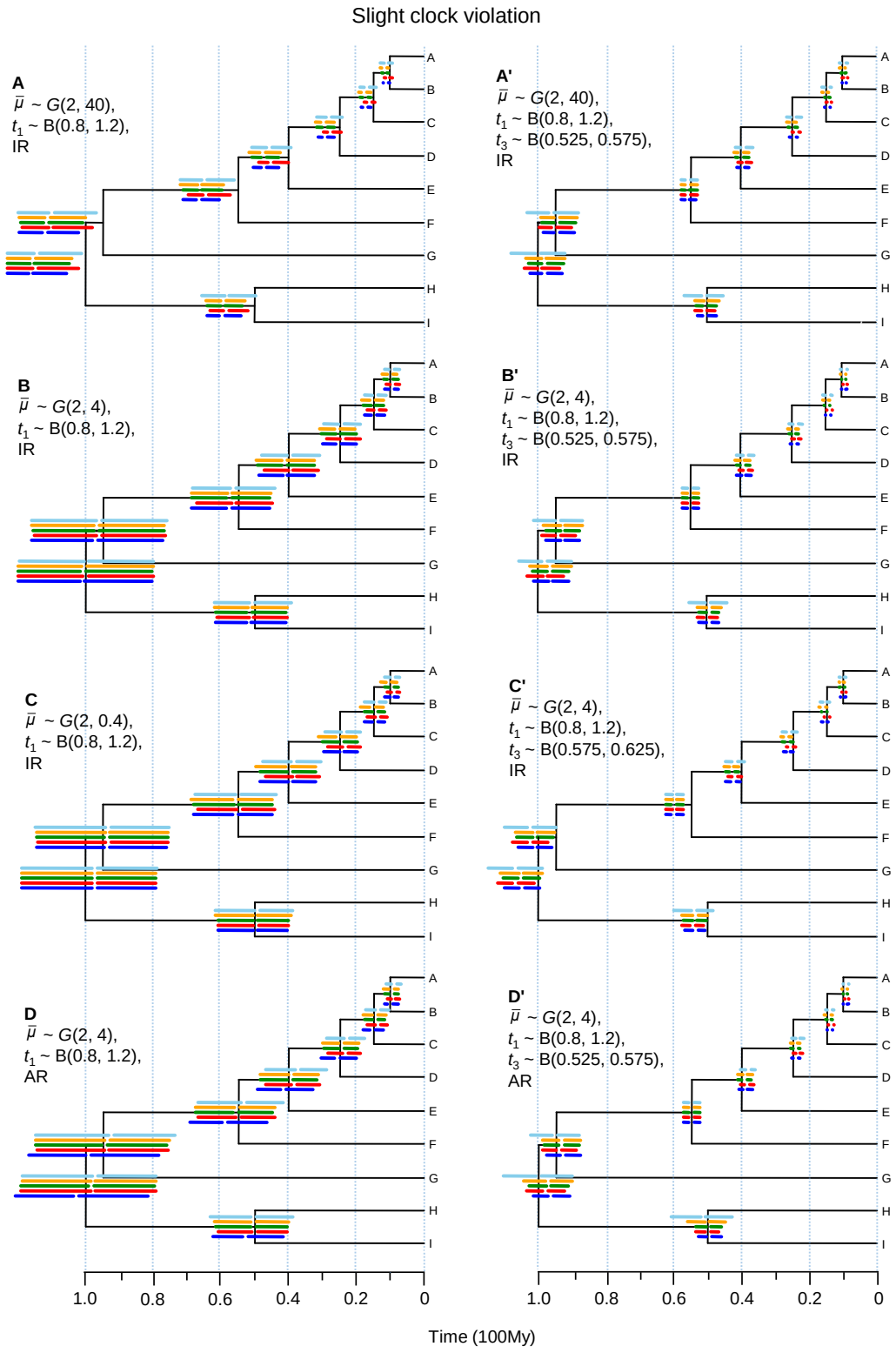

Figure 5.4: Posterior divergence time estimates from simulated data for each combination of rate prior, calibration strategy and rate-drift model, when the clock is slightly violated. See legend of Figure 5.3 for more details.

Table 5.3: Performance of different partitioning strategies in data simulated with slight clock violation.

Rate model	$\bar{\mu}$ ~	Calibration	rel. error					HPD width/age					$\sqrt{MSE}$					coverage probability (%)				
			C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)
IR	*G*(2, 40)	t_1 ~ B(0.8, 1.2)	0.148	0.187	0.179	**0.133**	0.176	0.29	0.22	0.20	0.22	**0.19**	0.083	0.092	0.091	**0.079**	0.091	**57**	0	1	39	1
	G(2, 4)	t_1 ~ B(0.8, 1.2)	**0.019**	0.031	0.026	0.029	0.026	0.43	0.42	0.41	**0.40**	0.41	0.054	0.054	0.053	**0.052**	0.053	100	100	100	100	100
	G(2, 0.4)	t_1 ~ B(0.8, 1.2)	0.032	0.029	**0.018**	0.043	0.023	0.43	0.44	0.41	**0.40**	0.41	0.054	0.055	**0.052**	0.053	**0.052**	100	100	100	100	100
	G(2, 40)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.525, 0.575)	0.019	**0.014**	**0.014**	0.031	0.019	0.17	0.12	**0.10**	**0.10**	**0.10**	0.022	0.016	**0.015**	0.017	**0.015**	**100**	**100**	**100**	74	96
	G(2, 4)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.525, 0.575)	0.027	**0.017**	0.023	0.037	0.029	0.16	0.12	**0.11**	**0.11**	**0.11**	0.022	0.019	**0.018**	0.019	0.019	**100**	99	97	67	88
	G(2, 0.4)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.525, 0.575)	0.034	**0.018**	0.024	0.038	0.031	0.18	0.12	**0.11**	**0.11**	**0.11**	0.025	**0.019**	**0.019**	**0.019**	**0.019**	**99**	**99**	97	67	87
	G(2, 4)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.575, 0.625)	0.063	0.075	0.067	**0.056**	0.060	0.17	0.13	**0.11**	**0.11**	**0.11**	0.041	0.038	0.036	0.037	**0.035**	**75**	35	32	44	42
AR	*G*(2, 4)	t_1 ~ B(0.8, 1.2)	0.051	**0.019**	0.021	0.039	0.042	0.45	0.43	0.41	**0.40**	0.41	0.058	0.053	**0.052**	**0.052**	0.056	100	100	100	100	100
	G(2, 4)	t_1 ~ B(0.8, 1.2) t_3 ~ B(0.525, 0.575)	0.033	**0.021**	0.024	0.037	0.028	0.17	0.14	**0.11**	**0.11**	**0.11**	0.026	0.020	**0.018**	0.019	0.019	90	**100**	95	67	89

Note.− See caption of Table 5.1.

Table 5.4: Performance of different partitioning strategies to estimate the ages of nodes 1 (top) and 4 (bottom) when the clock is slightly violated.

Rate model	$\bar{\mu}$ ~	Calibration	rel. error					HPD width/age					$\sqrt{MSE}$					coverage probability (%)				
			C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)	C (1P)	CP (2P)	PF (V)	G (50P)	GCP (100P)
IR	*G*(2, 40)	t_1 ~ B(0.8, 1.2)	**0.146**	0.157	0.160	0.150	0.164	0.22	0.19	**0.18**	0.21	**0.18**	**0.156**	0.164	0.167	0.159	0.170	**5**	0	0	0	0
			0.144	0.194	0.188	**0.136**	0.185	0.29	0.23	0.20	0.22	**0.19**	0.065	0.081	0.078	**0.059**	0.077	**82**	0	0	30	0
	G(2, 4)	t_1 ~ B(0.8, 1.2)	**0.002**	0.007	0.010	0.003	0.012	0.40	0.40	0.40	0.40	0.40	0.102	0.102	0.102	0.102	0.102	100	100	100	100	100
			0.014	0.035	0.033	**0.014**	0.029	0.43	0.42	0.42	**0.40**	0.41	0.044	0.046	0.045	**0.041**	0.044	100	100	100	100	100
	G(2, 0.4)	t_1 ~ B(0.8, 1.2)	0.027	**0.026**	**0.026**	0.027	**0.026**	0.40	0.40	0.40	0.40	0.40	0.105	0.105	0.105	0.105	0.105	100	100	100	100	100
			0.037	0.020	**0.013**	0.040	0.014	0.43	0.44	0.41	**0.40**	0.41	0.046	0.047	0.043	0.044	**0.042**	100	100	100	100	100
	G(2, 40)	t_1 ~ B(0.8, 1.2)	**0.009**	0.017	0.018	0.011	0.018	0.16	0.12	**0.10**	**0.10**	**0.10**	0.043	0.035	0.033	**0.029**	0.032	**100**	**100**	98	**100**	97
		t_3 ~ B(0.525, 0.575)	0.017	0.010	**0.009**	0.019	**0.009**	0.13	0.11	**0.10**	**0.10**	**0.10**	0.015	0.012	**0.011**	0.013	**0.011**	**100**	**100**	**100**	99	**100**
	G(2, 4)	t_1 ~ B(0.8, 1.2)	**0.020**	0.035	0.035	**0.020**	0.036	0.16	0.12	0.11	0.11	**0.10**	0.045	0.048	0.045	**0.035**	0.045	**100**	95	91	**100**	90
		t_3 ~ B(0.525, 0.575)	0.033	**0.013**	0.015	0.033	0.021	0.13	0.11	**0.10**	**0.10**	**0.10**	0.019	**0.013**	**0.013**	0.017	0.014	**100**	**100**	**100**	92	**100**
	G(2, 0.4)	t_1 ~ B(0.8, 1.2)	**0.022**	0.038	0.037	**0.022**	0.038	0.16	0.12	0.11	0.11	0.10	0.048	0.049	0.046	**0.036**	0.047	**100**	95	91	99	88
		t_3 ~ B(0.525, 0.575)	0.039	**0.014**	0.017	0.035	0.023	0.14	0.11	**0.10**	**0.10**	**0.10**	0.022	**0.013**	**0.013**	0.018	0.014	99	**100**	**100**	92	**100**
	G(2, 4)	t_1 ~ B(0.8, 1.2)	0.069	0.052	0.053	0.069	**0.051**	0.16	0.13	**0.11**	**0.11**	**0.11**	0.080	0.061	0.060	0.075	**0.058**	**79**	74	55	10	59
		t_3 ~ B(0.575, 0.625)	0.055	0.079	0.075	**0.054**	0.068	0.14	0.11	0.11	**0.10**	0.11	0.026	0.034	0.032	**0.024**	0.029	**83**	2	3	39	7
AR	*G*(2, 4)	t_1 ~ B(0.8, 1.2)	0.025	0.020	**0.019**	0.020	0.030	0.40	0.40	0.40	0.40	**0.39**	0.104	**0.103**	**0.103**	**0.103**	0.105	100	100	100	100	100
			0.067	0.018	**0.017**	0.035	0.047	0.45	0.43	0.41	**0.40**	0.41	0.053	0.045	**0.043**	**0.043**	0.046	100	100	100	100	100
	G(2, 4)	t_1 ~ B(0.8, 1.2)	**0.009**	0.025	0.028	0.019	0.035	0.20	0.14	**0.11**	**0.11**	**0.11**	0.053	0.046	0.042	**0.035**	0.045	**100**	99	97	**100**	91
		t_3 ~ B(0.525, 0.575)	0.041	0.020	**0.019**	0.034	0.021	0.12	0.11	**0.10**	**0.10**	**0.10**	0.021	0.014	**0.013**	0.017	0.014	91	**100**	**100**	93	99

Note.− See caption of Table 5.1.

5.3.3 Divergence times of plants

We estimated the divergence times of fifteen plant species using the five partitioning schemes and the tree topology of Figure 5.2. The PartitionFinder program generated a best-fitting scheme with 11 partitions. The posterior means and 95% HPD intervals of divergence times of the plant phylogeny are shown on Table 5.5. The time estimates were very similar for the three rate priors G(1, 100), G(1, 10) and G(1, 1) and thus only the estimates under the $\bar{\mu} \sim G(1, 10)$ prior are reported.

Figure 5.5 shows the divergence times and their HPD intervals estimated with the five partitioning schemes under the independent and autocorrelated rates models. The differences in time estimates among the partitioning schemes are large, even for some calibrated nodes. Under the independent rates model the estimated ages of the deep nodes (i.e. nodes 16, 17, and 29) become older as the number of partitions increase, whereas those of the other nodes become younger. For example, the age of pteridophytes (node 29) varies between 264 Ma (C scheme) and 368 Ma (GCP scheme) while the age of angiosperms (node 19) varies between 127 Ma (GCP scheme) and 204 Ma (C scheme) (Table 5.5). The time estimates for the angiosperms are within the minimum and maximum calibration bounds with the youngest estimate to be very close to the minimum bound (124 Ma). However, for node 28 the time estimates vary from 13 Ma (GCP scheme) to 70 Ma (C scheme) with the estimates under the G and GCP schemes to be well below the minimum bound (65 Ma).

The estimates among partitioning schemes using the autocorrelated rates model show similar high discrepancies. For example, the age of the root varies from 438 Ma (C scheme) to 453 Ma (GCP scheme) and the age of node 29 from 303 Ma (C scheme) to 375 Ma (PF, G schemes). The age estimates of the deepest nodes become older as the number of partitions increases. The time estimates with the autocorrelated rates model are in general older than those with the independent rates model. For example, the posterior mean of node 29 is 264 Ma and 296 Ma with the schemes C and CP, respectively, under the independent rates model, compared with 303 Ma and 347 Ma, under the autocorrelated rates model (Table 5.5).

In general the differences among the partitioning schemes are large, as was the case in the simulation analysis with severe clock violation when an incorrect rate-drift model was used. The highly partitioned schemes G and GCP tend to produce precise estimates, far from these of the other three schemes. In some cases those estimates are outside the calibration bounds (e.g. node 28) irrespective of the clock model, raising concerns about their accuracy.

Note that the estimate for the rate drift parameter σ^2 using the C scheme and the independent rates model was 0.58, indicating severe clock violation.

5.4 Discussion

Partitioning is a commonly used approach to account for variation in the substitution patterns among the sites of a molecular alignment in any phylogenetic analysis (Nylander, et al. 2004; Brandley, et al. 2005; Brown and Lemmon 2007). Although its importance has long been recognised (Yang 1996b) its effect on estimation of species divergence times has not been studied in detail.

Results from our simulation and real data analyses suggest that low partitioned schemes (e.g. C, CP schemes) produce uncertain time estimates (wide HPD intervals) but have high coverage probabilities. The use of finer partitions (e.g. PF, G, GCP schemes) increases precision but may lead to seriously biased estimates (low accuracy), especially when prior assumptions are incorrect. The number and quality of calibrations are very important. Multiple fossil calibrations increase accuracy, improve precision and reduce the differences in time estimates among partitioning schemes. However, incorrect calibrations exert a significant effect in time estimates and introduce bias no matter the partitioning scheme used. Results also indicate that the use of automated tools, such as PartitionFinder, to select the best-fitting partitioning scheme does not seem to always improve the inference.

The choice of the partitioning scheme seems to be more important when the clock is seriously violated. In that case a highly partitioned scheme in combination with incorrect prior assumptions such as an incorrect clock model may produce significant bias. Results from simulations showed that when the incorrect autocorrelated rates model was used, increasing the number of partitions produced higher bias with its direction to depend on the configuration and precision of fossil calibrations on the tree (Figure 5.3D). The use of many calibrations seems to improve accuracy but large biases may still remain for highly partitioned schemes (Figure 5.3D'). Thus the partitioning scheme may have a strong effect in divergence time estimation when the clock is seriously violated and attention should be given by performing a robustness analysis (e.g. see dos Reis, et al. 2015). Some previous studies have failed to identify differences in time estimates among partitioning schemes because they used many calibrations and focused on the analyses of closely related species where the clock approximately holds (Poux, et al. 2008; Voloch and Schrago 2012). However, when we analyzed the plant data set, where substantial rate variation was detected, we found large differences in the time estimates among partitioning schemes, irrespective of

Table 5.5: Posterior estimates of divergences times of plants (Ma) using different partitioning schemes

Rate drift model	Node	Clade	Prior		C (1P)		CP (2P)		PF (11)		G (78P)		GCP (156P)	
			Mean	(95%CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)
IR	16	Root	437	(417, 455)	440	(419, 456)	442	(421, 456)	451	(440, 457)	453	(448, 457)	454	(450, 458)
	17	Helianthus / Psilotum	412	(386, 443)	420	(389, 448)	427	(397, 452)	445	(431, 455)	452	(446, 456)	453	(449, 457)
	18	Angiosperms / Ginkgo	337	(306, 367)	332	(305, 363)	329	(304, 361)	326	(305, 349)	310	(302, 319)	305	(297, 311)
	19	Angiosperms	186	(124, 249)	204	(155, 251)	197	(155, 245)	184	(164, 206)	144	(136, 152)	127	(122, 133)
	20	Helianthus / Nymphaea	172	(111, 237)	189	(142, 237)	183	(140, 223)	172	(152, 191)	135	(128, 143)	120	(114, 125)
	21	Helianthus / Acorus	158	(102, 223)	166	(123, 211)	159	(122, 197)	147	(131, 164)	115	(109, 121)	101	(97, 106)
	22	Eudicots	134	(76, 211)	138	(97, 180)	131	(99, 167)	121	(106, 137)	95	(89, 101)	84	(80, 88)
	23	Helianthus / Eucalyptus	111	(38, 180)	109	(70, 148)	103	(74, 134)	93	(81, 107)	75	(70, 80)	67	(63, 70)
	24	Helianthus / Cornus	71	(0, 132)	82	(39, 120)	79	(47, 108)	74	(63, 87)	59	(54, 63)	53	(50, 57)
	25	Oxalis / Eucalyptus	71	(0, 131)	85	(44, 123)	81	(52, 113)	76	(64, 88)	63	(58, 67)	57	(53, 60)
	26	Monocots	130	(78, 192)	141	(102, 184)	137	(103, 171)	129	(114, 144)	96	(90, 103)	84	(80, 89)
	27	Yucca / Chamaedorea	101	(66, 152)	106	(77, 141)	103	(80, 130)	98	(88, 110)	59	(53, 64)	51	(48, 56)
	28	Elaeis / Chamaedorea	74	(65, 81)	70	(64, 80)	68	(64, 76)	65	(62, 67)	17	(15, 20)	13	(12, 15)
	29	Ferns	146	(0, 369)	264	(138, 386)	296	(185, 385)	339	(303, 374)	362	(349, 376)	368	(356, 378)
AR	16	Root			438	(418, 455)	439	(418, 455)	443	(426, 456)	452	(446, 457)	453	(446, 458)
	17	Helianthus / Psilotum			416	(387, 443)	419	(390, 446)	433	(414, 452)	450	(443, 455)	450	(443, 455)
	18	Angiosperms / Ginkgo			342	(313, 368)	347	(320, 368)	361	(348, 371)	358	(347, 367)	344	(332, 355)
	19	Angiosperms			229	(197, 254)	232	(204, 253)	237	(222, 251)	191	(178, 204)	166	(155, 176)
	20	Helianthus / Nymphaea			219	(187, 245)	221	(195, 243)	226	(210, 240)	179	(167, 192)	156	(146, 166)
	21	Helianthus / Acorus			194	(164, 221)	196	(170, 219)	199	(184, 213)	152	(141, 164)	131	(122, 140)
	22	Eudicots			162	(129, 192)	165	(139, 191)	168	(153, 183)	122	(113, 132)	106	(99, 114)
	23	Helianthus / Eucalyptus			119	(87, 154)	122	(94, 152)	127	(112, 143)	90	(82, 98)	80	(74, 87)
	24	Helianthus / Cornus			95	(68, 130)	100	(73, 127)	106	(91, 121)	73	(66, 80)	66	(60, 71)
	25	Oxalis / Eucalyptus			99	(69, 131)	102	(74, 129)	108	(93, 124)	76	(69, 83)	68	(63, 74)
	26	Monocots			177	(149, 205)	179	(155, 202)	181	(167, 194)	133	(122, 143)	114	(105, 122)
	27	Yucca / Chamaedorea			141	(116, 167)	142	(122, 165)	143	(131, 155)	92	(83, 100)	77	(70, 84)
	28	Elaeis / Chamaedorea			71	(64, 80)	69	(64, 77)	65	(63, 68)	29	(24, 33)	21	(18, 24)
	29	Ferns			303	(153, 396)	347	(288, 397)	375	(353, 396)	375	(362, 387)	369	(357, 380)

Note.−Node numbers are according to Figure 5.2. The rate prior was $\bar{\mu}$ ~ G(1, 10). See caption of Table 5.1 for more details.

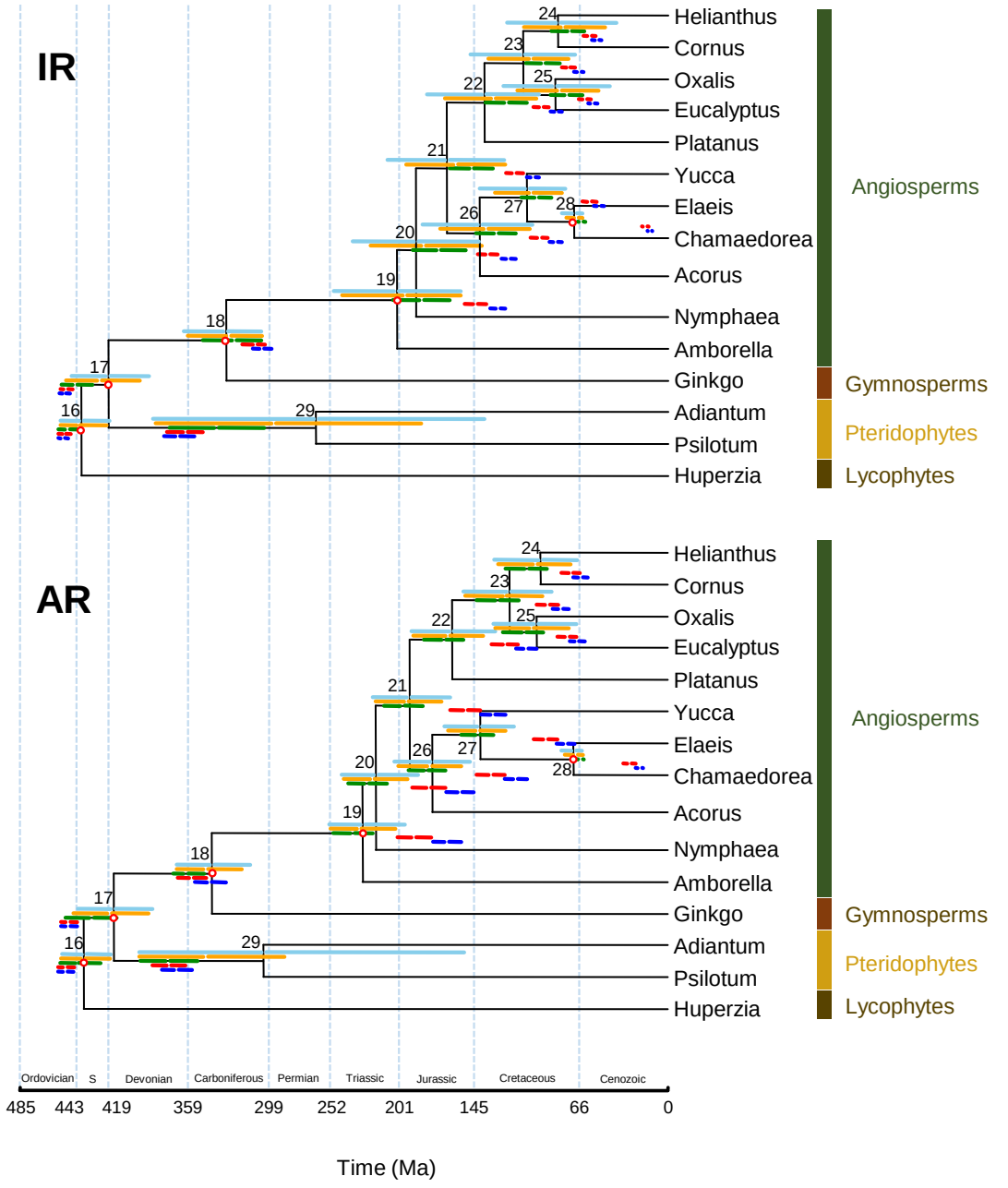

Figure 5.5: Posterior divergence times of fifteen plant species using five partitioning schemes and two rate drift models. Horizontal bars represent the 95% high posterior density intervals under the five partitioning strategies with the gaps to denote the posterior mean. These are (from the top to the bottom): (i) concatenation (C), 1 single partition (light blue); (ii) codon positions (CP), 2 partitions (codon positions 1+2 vs 3) (yellow); (iii) PartitionFinder (PF), 11 partitions (green); (iv) gene (G), 78 partitions (red); and (v) both gene and codon positions (GCP), 156 partitions (blue). The timetrees shown in black were estimated using the C scheme. Node numbers are reported and calibrated nodes are indicated by red circles. IR: independent-rates model, AR: autocorrelated-rates model, S: Silurian.

the rate drift model used. Especially the use of the highly partitioned schemes G and GCP gave very different estimates than all other schemes, which in some cases were outside the calibration bounds raising concerns on their accuracy.

In the simulation analysis when we use a single calibration in the root, correct rate prior and rate drift model, increasing the number of partitions further from the CP scheme does not reduce further the HPD widths (Figure 5.3B and B') as one might expected (Zhu, et al. 2015). This is probably because the alignment is long (75,000 bp in total) and also the increase in the number of partitions is not based on the addition of new molecular data but on the division of the whole alignment in finer partitions. In the simulation analysis we also saw that the time estimates under the slow rate prior were more biased (overestimates) than under the fast rate prior (underestimates) for all schemes, although the two priors were wrong by the same magnitude (10 times slower and 10 time times faster, respectively). This could be because under a fast rate prior time estimates are expected to be younger, but node ages are bounded below by 0 (the age of the youngest node can't be lower than 0). Under a slow rate prior time estimates are expected to be older with no bound to be applied to the root age, letting them to move back with any constraint.

Many of previous studies have used protein-coding genes to estimate species divergence times (Meusemann, et al. 2010; dos Reis, et al. 2012; Misof, et al. 2014). We thus simulated gene alignments from a phylogeny of nine species using a codon site model which allows for different ω ratios across sites to mimic such data sets. Although a branch-site model allowing for additional variation among branches in the ω ratio would be more realistic, that would increase the complexity of the simulation process with probably only minimal effects on divergence time estimation. Another important aspect of the simulation analysis is the use of equal codon frequencies in all genes. This is an unrealistic assumption and variable codon frequencies among genes may affect divergence times estimation. In such a case, a partitioned analysis might be more useful as it accounts for heterogeneity in base frequencies across the alignment, but further research is needed. Moreover, the same gene tree was used to simulate gene alignments, which is unrealistic (Nichols 2001). The use of different gene trees is not expected to highlight differences among partitioning schemes because in the Bayesian dating analysis with the MCMCTREE program the same phylogeny is assumed for all partitions.

6 Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales

In the previous chapter we evaluated the performance of different partitioning strategies in Bayesian estimation of species divergence times. In this chapter we will use the Bayesian algorithm implemented in the MCMCTREE program to estimate the divergence times of 54 metazoan species. The effect of incomplete lineage sorting was found to be less important in old phylogenies (see chapter 4); thus it is ignored in the following analysis.

6.1 Introduction

Estimating the timing and rate of animal evolution has been one of the most appealing and enduring problems in evolutionary biology. The knowledge of early metazoan diversifications may provide an insight into the underlying processes of animal evolution. The fossil record suggests a possible divergence of metazoans before 635 Ma, during the Cryogenian (Love, et al. 2009; Maloof, Rose, et al. 2010). This is further supported by molecular clock dating studies (Sperling, et al. 2010; Erwin, et al. 2011). Despite the consistency on the Cryogenian origin of the crown Metazoa, the evidence for the diversification of Bilateria remains controversial. Molecular dating studies place the origin of Bilateria during the Ediacaran period (635−541 Ma) (Peterson, et al. 2008; Erwin, et al. 2011) but the fossil record suggests a massive radiation of Bilateria phyla after the first 20 My of the Cambrian (541−485 Ma) with no unequivocal records of crown bilaterians prior to the Cambrian (Budd 2008; Maloof, Porter, et al. 2010; Erwin, et al. 2011). Nevertheless, there is increasing acceptance of a Precambrian history to animal evolution and only its extend remains open to question. Was there a rapid radiation of crown Bilateria close to the base of Cambrian (Budd 2008; Lee, et al. 2013)? Or is there an extensive Precambrian history which extends into Cryogenian and the absence of a fossil record simply reflects preservation and/or interpretation biases?

A promising approach to deal with the issue has been to estimate the timescale of animal evolution using molecular clock methodology. Indeed, several efforts have been made but the estimated times have been inconsistent among studies. For example, the age of Bilateria has been estimated at 700 Ma (Peterson and Butterfield 2005) and 573 Ma

(Peterson, et al. 2004), both before the Ediacaran−Cambrian boundary. The disparity between molecular clock estimates and clade ages suggested by the fossil record have been diminished with the increase of molecular data and the improvement of molecular clock methodology, especially concerning the accommodation of rate variation. More recent molecular dating studies have been performed within the Bayesian framework because it provides a straightforward way to integrate much of the uncertainty associated with divergence time estimation such as uncertainties in fossil information, due to rate variation among lineages (the relaxed clock), in branch length estimation due to limited molecular data, in the phylogenetic relationships of species under study (the tree topology) and parameters such as data partitioning. However, the time estimates are still largely inconsistent among methods and parameter settings (Peterson, et al. 2008; Erwin, et al. 2011; Lee, et al. 2013; Rota-Stabelli, et al. 2013). Moreover, the cumulative impact of those uncertainties on the precision of evolutionary timescales has not been studied in detail.

Here, we use a Bayesian method to estimate the divergence times of Metazoa and show that the precision of molecular clock time estimates has been grossly overestimated. We perform a sensitivity analysis and explore in detail the impact of different sources of uncertainties in posterior time estimates. We use a large amino acid alignment (38,557 sites) of 203 nuclear coding genes from 54 species in combination to 34 fossil calibrations. We use four fossil calibration strategies which reflect different interpretations of the fossil record and, show that these have a dramatic impact on estimated times. We also explore the use of different relaxed clock models and show that the molecular clock is significantly violated at this level of divergence. We test for the effects of different data partitioning strategies and show that this, too, has a significant impact on divergence time estimates. Finally, we show that competing phylogenetic hypotheses lead to different divergence time estimates. The estimated evolutionary timescale accommodating these uncertainties has low precision preventing the inference on plausible scenarios for the emergence and evolution of early animal life forms. Although some of this uncertainty can be reduced by using more molecular data and reducing the topological uncertainty, the limitations of the fossil record and the confounding effect of times and rates will remain, hampering out efforts to draw conclusions on the timeline of metazoan diversification.

6.2 Methods

6.2.1 Molecular data and tree topology

Two independent molecular data sets from Philippe et al. (2011) and Erwin et al. (2011) were combined into a single amino acid alignment. Missing or incomplete proteins in the original alignments were updated with the non-redundant protein database from GenBank. In addition, 5 new species (*Homo sapiens*, *Mus musculus*, *Ornithorhynchus anatinus*, *Tribolium castaneum* and *Caenorhabditis elegans*) were added in order to accommodate more calibration points. For each gene, amino acid sequences of all species were aligned with PRANK (Loytynoja and Goldman 2005) and the alignment gaps were removed using GBLOCKS (Castresana 2000). The combined alignment consists of 203 nuclear proteins (38,577 amino acid positions) from 71 species (missing data 21.49%).

As the relationships among many taxa remain unresolved, 17 species were removed from the dataset to reduce the uncertainty in the topology. This resulted in a smaller alignment of the remaining 54 species (missing data 13.97%). The tree topology for these 54 species has 5 uncertain nodes; four of them can be rearranged in three ways and one of them can be rearranged in two ways, giving $3^4 \times 2 = 162$ possible fully resolved trees which were analysed. One of those trees (Figure 6.6), mainly based on Philippe et al. (2011) was chosen for the main analysis while the other 161 trees were used to assess the robustness of the time estimates to the various topologies.

6.2.2 Data partitioning

Six partitioning schemes were considered. First, we used the relative rates to partition the combined alignment. Amino acid distance estimates for each gene were obtained from pairwise comparisons between *Strongylocentrotus purpuratus* and *Hydra magnipapillata* under the WAG+Γ_4+F model in CODEML (Yang 2007). These two species were chosen because of their deep divergence time and because they have the most complete sequence data. For one missing gene of *Strongylocentrotus purpuratus*, the same gene of *Saccoglossus kowalevskii*, its close relative, was used instead. Assuming that the divergence time is the same for all genes, the estimated distances reflect the relative evolutionary rates. These distances were used to assign the 203 genes into two, four, five and ten, rate categories (partitions), in addition to the single partition, thus forming five partitioning schemes.

There is a possibility, however, that the rates estimated from these two chosen species may not be representative of the rates across branches. To address this issue, the suitability of the partitioning strategy was assessed by calculating the branch lengths of each partition in each partitioning scheme using the WAG+Γ_4+F model. If the use of the pairwise distances is suitable in partitioning the data, the sum of the branch lengths (i.e. tree length) is expected to be approximately ordered from a partition with the lowest rate category to one with the highest rate category. This was found to be the case.

Second, the data were divided into two partitions according to hydrophobicity, using the hydropathy index (Kyte and Doolittle 1982). An average of the hydropathy index for each site in the alignment was calculated (gaps excluded). Then the site was classified as hydrophilic if the averaged hydropathy index was negative, otherwise it was classified as hydrophobic. The time estimates from this partitioning scheme were very similar to the two partitions scheme according to rate and thus are not reported.

6.2.3 Fossil calibrations

We constrained the ages of thirty-four nodes in the metazoan tree based on fossil information from Benton et al. (2009) with updates from Benton et al. (2015) and Warnock et al. (2012). The minimum ages were determined from the oldest uncontroversial record belonging to one of the two sister clades. The maximum ages were derived from the base of the youngest stratigraphic range or geological formation known not to contain any members of the clade of interest (Benton and Donoghue 2007; Donoghue and Benton 2007). A critical fossil is the Ediacaran *Kimberella* (552.85 Ma) which we interpret as a protostome, thus providing the minimum age constraint for Metazoa, Eumetazoa, Bilateria and Protostomia.

We translated the fossil calibrations into statistical distributions mapped onto the nodes of the metazoan tree (see §2.3). Four calibration strategies (Table D.1) were used to assess robustness of time estimates to the calibration choice.

1) *Strategy 1* (S1): The 34 calibrations are represented as uniform distributions between the minimum and maximum bounds. Bounds are soft, and we assigned 0.1% and 2.5% tail probabilities that minimum and maximum bounds are violated (but we used 0.1% for both bounds on the age of the root). A variation of the S1 was also tested where the Cambrian snail *Aldanella* (532 Ma) was used instead of *Kimberella* to constrain the basal clades of Metazoa, Eumetazoa, Bilateria and Protostomia. This change did not affect the results.

2) *Strategy 2* (S2): 13 calibrations are represented as skewed-normal distributions. This was done for nodes for which the oldest in-group fossil is thought to be very close to the actual parent node being calibrated. The parameters of the skew-normal (location, scale,

shape) were chosen to provide a distribution with the mode near the minimum bound and the tail extending towards old ages, with the 0.3% and 97.5% quantiles of the distribution lying roughly at the equivalent minimum and maximum bounds from S1. These calibrations represent an optimistic interpretation of the fossil minima as a close approximation of the true clade age. The remaining 21 nodes are as in S1.

3) *Strategy 3* (S3): The same 13 nodes are calibrated using a truncated Cauchy distribution (Inoue, et al. 2010) with 0.1% left tail probability, with the mode of the distribution on the minimum bound, and with tail parameter equal to 10, leading to a long right tail for the distribution. No maximum bound is imposed on these nodes. The root node has an older minimum bound (634.9 Ma) accounting for alternative fossil interpretations.

4) *Strategy 4* (S4): Like S3, but the tail parameter is 0.1 rather than 10, producing a truncated-Cauchy calibration with a much shorter tail.

Note that the Cauchy is a heavy-tailed distribution, that is, it places considerable probability mass on its tail in contrast to the skew-normal (in S2) which is light-tailed. Strategies 3 and 4 represent a pessimistic interpretation of palaeontological evidence in which the first fossil records of clades are a poor approximation of their antiquity. Moreover, under strategies 1 and 2, the age of crown Metazoa has the minimum constraint based on a protostome interpretation of the Ediacaran *Kimberella*, whereas in strategies 3 and 4 it is based on the disputed biogeochemical evidence of Cryogenian demosponges (Love, et al. 2009; Antcliffe, et al. 2014).

6.2.4 Divergence time estimation

All molecular dating analyses were performed using the program MCMCTREE v4.8 (Yang 2007). The time unit was set to 100 My. The prior on times was constructed using the fossil calibrations and the birth-death process with parameters $\lambda = \mu = 1$, $\rho = 0$, which specify a diffuse uniform kernel and hence a diffuse prior. MCMCTREE applies a truncation to the user-specified densities to ensure that ancestral nodes are older than descendant nodes. This may result in marginal priors very different from the specified calibration densities (see §2.3). To assess the effect of truncation, the marginal priors were obtained by running the MCMC without sequence data and were compared with the calibration densities. In addition, comparing the marginal priors with the marginal posteriors allows the relative impact of the prior and the sequence information to be assessed. The marginal priors for all the nodes are shown in Figure D.1 and summarized in Table D.2.

Because the molecular alignment is large, the likelihood was calculated approximately to save computational time (see §2.3.2). CODEML was used to estimate the branch lengths.

The LG+Γ_4+F amino acid substitution model was used in all partitioning schemes. For the combined alignment and the two partitions scheme according to hydrophobicity we also used the GTR+Γ_4+F model to assess robustness of the time estimates to the substitution model. The results were very similar to those under the LG+Γ_4+F model and thus are not reported here.

Both the independent-rates and the autocorrelated-rates models were used. The gammaDirichlet prior was used for both model parameters (see §2.3.4). The prior on the mean rate (or the ancestral rate) was set to $G(2, 40)$. This is a diffuse prior with mean 0.05, meaning 5×10^{-10} amino acid substitutions per site per year. The overall mean was derived from the average pairwise amino acid distances between the 203 proteins of *Hydra magnipapillata* and *Strongylocentrotus purpuratus* (0.29 substitutions/site) assuming a divergence time of 636.1 Ma, so that the mean rate is $0.29/6.361 = 0.046 \approx 0.05$. The prior for σ^2 (or v) was set to $G(1, 10)$, with mean 0.1, indicating serious violation of the clock.

The number of iterations, the burn-in and the sampling frequency were adjusted in test runs of the program. The step sizes of the proposals used in MCMC were adjusted such that the acceptance proportions were close to 0.3. Convergence was assessed by comparing the posterior means from two independent runs with different starting values. The resulting posterior distribution from one of the two runs was summarized as the means and 95% HPD intervals.

6.3 Estimates of metazoan divergence times

6.3.1 The impact of fossil calibrations on divergence time estimates

To assess the robustness of estimated Metazoan divergences to calibration choice, we established temporal constraints on the ages of 34 nodes of the Meatazoan phylogeny based on fossil evidence (Table 6.1). These were then translated into probability densities according to four calibration strategies, reflecting different interpretations of the fossil evidence (Table D.1). The program MCMCTREE was used to obtain posterior time estimates under these four strategies on the fixed tree topology of Figure 6.6. All gene alignments were concatenated and analysed as a single partition under the LG+Γ amino acid substitution model and the independent-rates model for among branches rate variation. In all instances, we first ran the analyses without sequence data to establish the effective time prior and evaluate the impact of truncation (Inoue, et al. 2010; Warnock, et al. 2012).

Calibration strategy has a large impact on estimated divergence times (Figure 6.1, Table D.3, Figure D.2). Specifically, when the skew-normal distribution is employed (strategy 2), the resulting posterior time estimates agree largely with those obtained using a uniform prior time distribution (strategy 1; Table 6.1, Figure 6.2A). In contrast, calibration densities modelled with the Cauchy distribution (strategies 3 and 4) exhibit strong truncation effects in the time priors (Figure 6.1C), resulting in substantially older time estimates (Figure 6.1D). This can be seen, for example, in association with Bilateria, Deuterostomia and Protostomia, where truncation caused the effective priors to place considerable probability mass beyond the maximum bound of 636.1 Ma (Figure 6.1C). This differs significantly from the specified calibration densities (Figure 6.1B), resulting in posterior time estimates that are substantially older than those derived using strategies 1 and 2 (Figure 6.1D). Thus truncation can have dramatic and perhaps surprising effects. These effects may be hard to predict, highlighting the challenges in constructing fossil calibrations, as calibrations based on the same fossil information can unintentionally lead to dramatically different estimates of divergence times.

Age estimates for the younger nodes are similar under all four calibration strategies (e.g. nodes 68, 86, 92; Table D.3). However, the posterior age estimates of nodes close to the root exhibit dramatic differences among the different calibration strategies. This is probably because of the paucity of palaeontological evidence close to the root and more severe truncation effects. Strategies 3 and 4 yield timescales that strongly favour an early Cryogenian (834-780 Ma) diversification, evidently constrained by the root age, while the age estimates arising from calibration strategies 1 and 2 are compatible with metazoans diversifying at any time within the Cryogenian, though these analyses are not otherwise very informative (Figure 6.1D).

Calibration strategies 1 and 2 are based on a protostome interpretation of the Ediacaran *Kimberella* (552.85 Ma), to constrain the minimum time of divergence of Protostomia, Bilateria, Eumetazoa and Metazoa (Table 6.1). However, to some, there is no unequivocal fossil evidence of metazoans prior to the Cambrian. In this view, interpreting *Kimberella* as a protostome leads to unduly ancient time estimates. To assess the impact of using *Kimberella* as a minimum constraint on the age of the protostome clade, we employed a variation of calibration strategy 1 in which the next-oldest record of Protostomia and oldest unequivocal total-group mollusc, the Cambrian *Aldanella yanjiahensis* (532 Ma), was used in place of *Kimberella*. The resulting divergence time estimates are effectively the same as those derived using strategy 1 (Figure D.4). Thus, even under the assumption that the fossil record of Metazoa is limited to the Cambrian, the results suggest an Ediacaran origin for most crown bilaterian phyla, a late Cryogenian - early Ediaracan origin of crown-Bilateria, and early Cryogenian origin of crown-Metazoa.

Table 6.1: Minimum and maximum fossil constraints and 95% HPD of posterior divergence times (Ma) for various metazoan clades.

Node	Crown group	Min.	Max.	S1, IR, 1P		S2, IR, 1P		S1, IR, 10P		S1, AR, 1P		Composite	
55	Metazoa	552.85	833	680.6	832.7	716.2	833.4	786.8	833.5	649.8	763.9	649.8	833.5
58	Eumetazoa	552.85	636.1	630.7	652.9	649.5	714.2	712.2	746.2	625.9	648.0	625.9	746.2
59	Cnidaria	529	636.1	533.3	620.5	537.7	631.9	596.2	641.7	587.4	629.0	531.5	641.8
63	Bilateria	552.85	636.1	615.1	637.8	624.2	672.3	665.6	688.3	595.7	618.7	595.7	688.3
64	Deuterostomia	515.5	636.1	593.7	627.9	598.0	649.6	639.5	662.3	587.2	610.6	587.2	662.3
65	Chordata	514	636.1	555.4	611.3	558.1	622.2	609.0	635.7	573.9	600.6	555.4	635.7
66	Olfactores	514	636.1	516.6	583.6	524.3	588.0	568.0	600.0	551.2	587.0	516.3	600.0
68	Vertebrata	457.5	636.1	459.6	527.9	467.1	527.6	483.3	512.9	481.4	533.8	459.3	533.8
69	Gnathostomata	420.7	468.4	432.9	468.7	433.9	468.6	436.2	451.3	440.5	468.9	432.1	468.1
70	Osteichthyes	420.7	453.7	420.6	444.1	420.6	443.9	420.6	425.0	420.6	438.1	420.6	444.2
71	Tetrapoda	337	351	338.3	351.4	338.4	351.5	346.5	352.1	345.8	352.2	338.2	354.0
72	Amniota	318	332.9	318.0	331.4	318.0	331.1	318.0	321.5	318.0	323.7	318.0	331.5
73	Mammalia	164.9	201.5	165.1	200.7	164.9	200.5	164.8	186.5	167.8	202.8	164.8	204.7
74	Euarchontoglires	61.6	164.6	61.4	140.2	61.4	135.3	61.3	67.3	61.6	124.7	61.2	140.3
75	Cyclostomata	358.5	636.1	358.1	458.0	358.1	455.8	358.3	416.5	378.1	494.3	358.0	494.3
76	Xenambulacraria	515.5	636.1	569.8	614.5	575.9	632.2	617.6	639.9	577.8	603.0	569.3	639.9
77	Ambulacraria	515.5	636.1	534.6	591.3	538.5	603.5	572.6	600.1	556.0	586.9	534.1	603.1
80	Hemichordata	504.5	636.1	504.2	537.6	504.2	540.0	504.1	511.4	504.2	535.8	504.1	540.0
82	Protostomia	552.85	636.1	598.0	626.4	603.6	647.5	635.3	653.5	578.1	599.0	578.1	653.1
85	Annelids-Molluscs	534	636.1	552.3	586.1	554.1	591.7	577.4	595.1	556.4	572.5	552.2	595.1
86	Capitellid-Polychete-leech	476.5	636.1	476.3	548.1	480.9	550.9	476.3	517.5	503.5	548.7	476.3	550.9
90	Mollusca	534	549	538.4	549.6	539.1	549.7	545.8	550.3	540.9	549.5	538.3	550.3
91	Bivalve-Gastropod	530	549	530.0	539.1	530.0	538.6	530.0	532.6	530.0	536.9	530.0	539.2
92	Gastropoda	470.2	549	470.0	508.3	470.3	506.2	470.0	478.8	470.5	512.6	470.0	512.6
96	Ecdysozoa	528.82	636.1	577.8	613.2	581.9	627.1	608.8	628.9	566.5	585.8	566.5	628.9
97	Nematoda-Arthropoda	528.82	636.1	561.4	599.8	563.8	608.3	589.8	610.4	557.2	575.5	557.2	610.4
98	Lobopodia	528.82	636.1	545.1	582.8	547.8	588.5	568.5	587.0	546.1	561.7	545.1	588.5
99	Euarthropoda	514	636.1	530.8	559.4	531.9	560.7	543.3	556.2	533.0	540.9	530.8	560.7
100	Mandibulata	514	531.22	523.4	532.3	524.0	532.3	530.3	536.1	528.1	532.8	523.4	536.1
101	Pancrustacea	514	531.22	514.0	522.8	514.0	522.3	514.0	517.5	514.0	517.6	514.0	522.8
102	Copepoda-Branchiopoda	499	531.22	499.0	510.1	498.9	509.2	498.9	500.5	499.0	506.4	498.9	510.1
105	Eumetabola	305.5	413.6	305.3	396.8	305.3	393.1	305.3	335.8	318.5	418.3	305.2	418.3
106	Pycnogonida-other chelicertates	497.5	531.22	497.5	526.1	497.5	525.8	497.4	509.1	497.5	518.5	497.4	526.3
107	Acari-Arenacea	416	531.22	415.9	479.9	415.8	477.5	415.8	436.4	419.6	492.5	415.7	492.5

Note: Posterior times are the 95% HPD interval, estimated with MCMCTree v4.8 under the LG+Γ_4+F model. S1: Calibration strategy 1. S2: Strategy 2. IR: Independent-rates model. AR: Autocorrelated-rates model. 1P: The 203 proteins analysed as a single partition. 10P: The proteins are grouped into 10 partitions according to their evolutionary rates. Node numbers as in Figure 6.6. Nodes in bold have calibrations that differ in S1 and S2. Composite: 95% HPD interval is a composite of the 95% HPD intervals across all the analyses, except those under S3 and S4 and under alternative topologies.

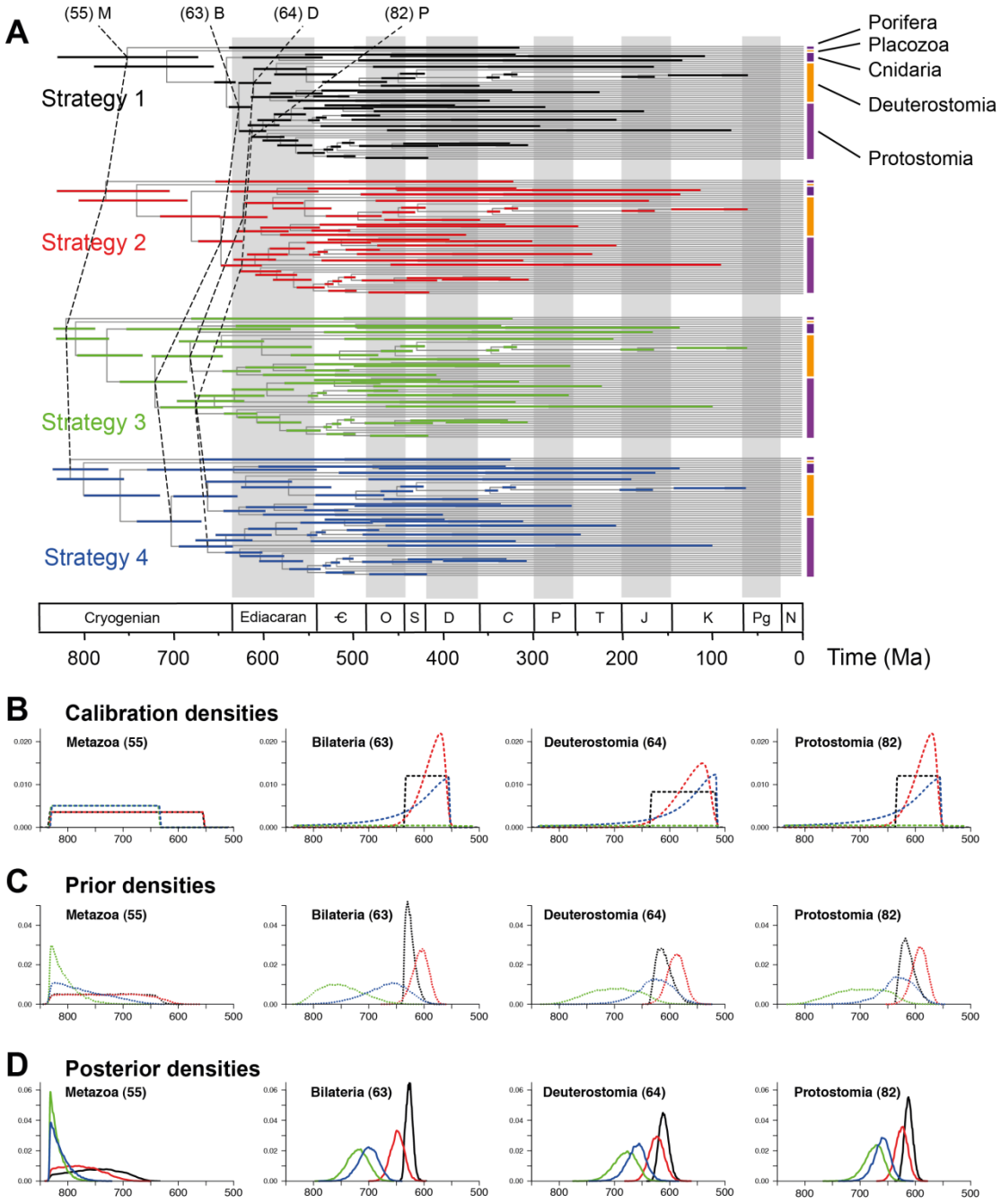

Figure 6.1: The effect of fossil calibrations on posterior divergence time estimates of metazoans. (A) Timetrees showing posterior divergence time estimates for major metazoan groups. Nodes are drawn at the posterior means obtained and horizontal bars represent 95% HPD intervals. Estimates were obtained with MCMCTREE using the LG+Γ_4+F model, independent-rates, and with the 203 proteins concatenated into a single alignment. Names of taxa are as in Figure 6.6. (B-D) Calibration, prior and posterior densities for four ancient nodes in the metazoan phylogeny. Node numbers (according to Figure 6.6) are in parentheses.

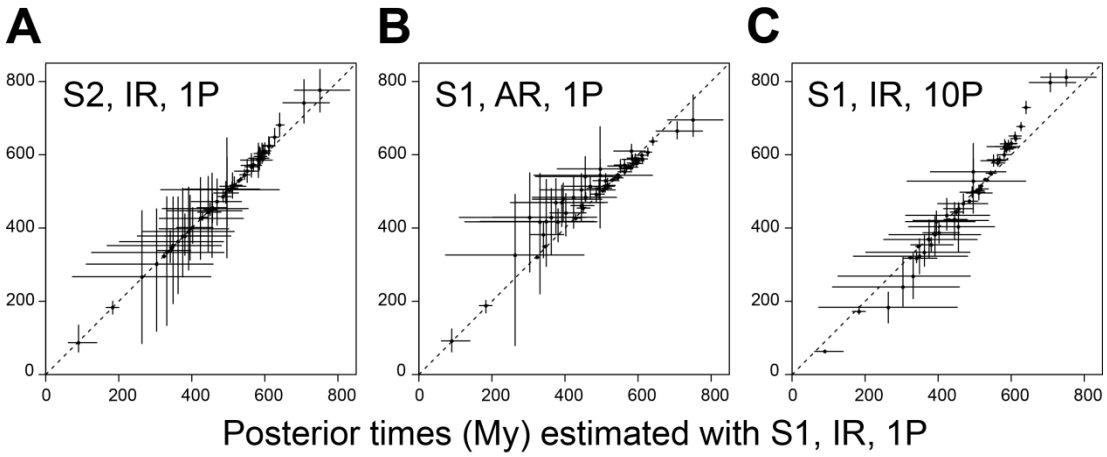

Figure 6.2: Sensitivity of time estimates to fossil calibrations, rate model and number of partitions. The posterior mean times estimated under calibration strategy 1, independent-rates and a single partition are plotted against: (A) estimates using strategy 2, (B) estimates under the autocorrelated-rates model, and (C) estimates obtained when the 203 gene alignments are divided into 10 partitions according to substitution rate. The bars are the 95% HPDs.

6.3.2 The impact of strong violation of the molecular clock in ancient timescales

When rate variation across a phylogeny is extreme (i.e. the clock is seriously violated), the rates calculated on parts of the phylogeny where fossil calibrations are available will serve as bad proxies to estimate divergence times in other parts of the tree. In such cases divergence time estimation is challenging and the analysis becomes sensitive to the rate model used.

To examine the impact of the rate model we re-estimated the divergence times of metazoans using the autocorrelated-rates model under calibration strategy 1. We found that the relaxed-clock model has a strong impact on the estimated divergences (Table 6.1, Figure 6.2B). The results show that many posterior time estimates for young nodes using the autocorrelated-rates model are older than those derived using the independent-rates model, whereas a few nodes, especially the deep ones, are younger (Table 6.1, Figure 6.2B). For example, the divergences of Metazoa (764–650 Ma), Bilateria (619–596 Ma), Deuterostomia (611–587 Ma) and Protostomia (599–578 Ma), are substantially younger.

The autocorrelated-rates model penalises extreme rate variation over short time intervals while allowing large rate variation among distant clades. This contrasts with the

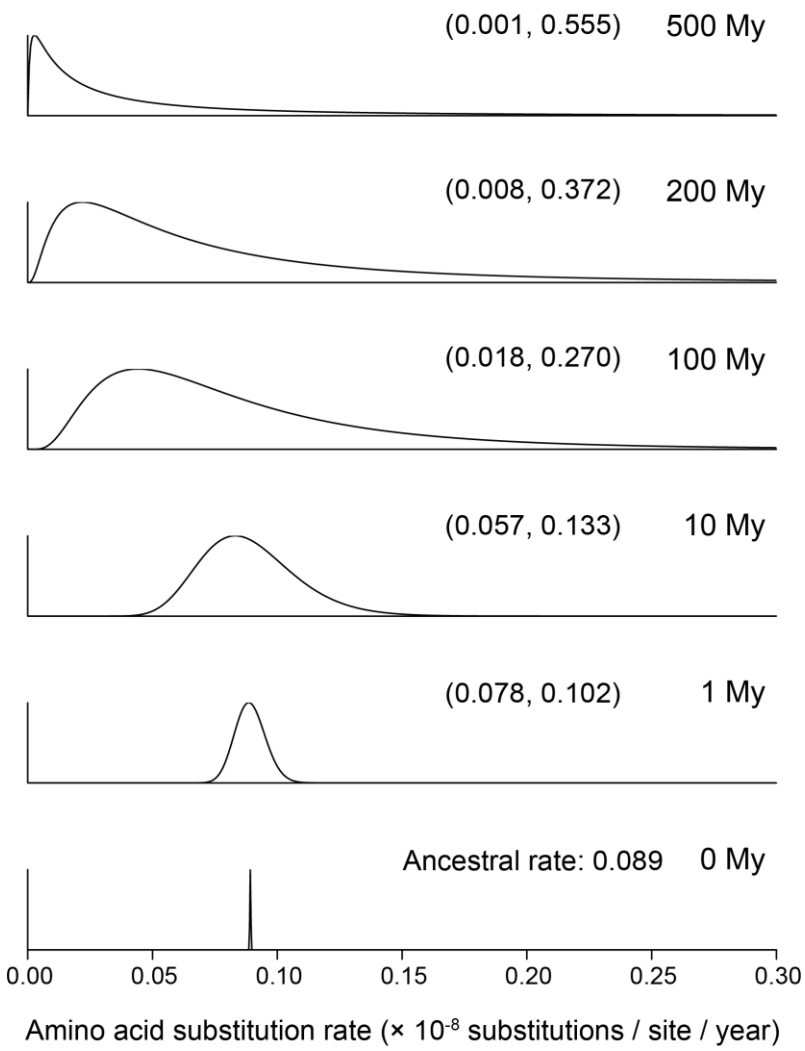

Figure 6.3: Explosive relaxation of molecular rates during Metazoan evolution.
In the autocorrelated-rates model (AR), the rates at the tips of a star phylogeny are log-normally distributed with mean r_A (the ancestral rate at the root) and log-variance of the rate $\sigma^2 = tv$. For the metazoan phylogeny, the posterior mean of r_A is 0.089 s/s/100My and of v is 0.468/100My. The graph shows the evolution of the rate of molecular evolution through 500 My of metazoan history assuming the AR model to be correct. The numbers in brackets are the 95% equal-tail range of the distribution of the rate for the given time. As the star phylogeny evolves, the variance of the rates increases exponentially. After 500My evolution, the 95% equal-tail range encompasses two orders of magnitude. Note that in case of the independent-rates model with μ = 0.089/100My and σ^2 = 0.468/100My, the shape of the log-normal distribution is the same as that for 100My for the autocorrelated-rates at any time point.

independent-rates model, which assumes that the variance of the rate is independent of the divergence time, so that the variance is the same whether the species are closely or distantly

related. Figure 6.3 shows the change in the shape of the log-normal distribution of rates under the autocorrelated-rates model across 500 My of evolution and highlights the extreme level of rate variation in the metazoan phylogeny. At short time scales, the distribution is more symmetrical and has a smaller variance than at longer time scales. In the case of the independent-rates model with μ = 0.089/100My and σ^2 = 0.468/100 My, the log-normal distribution has the same shape as that for 100 My for the autocorrelated-rates (Figure 6.3, 3rd plot).

6.3.3 The impact of data partitioning

Partitioning of the molecular sequence alignment may impact on divergence time estimates (Zhu, et al. 2015). To test whether the choice of partitioning scheme has an influence on time estimation the protein alignment was divided into two, four, five, and ten partitions, according to the relative substitution rates among genes. The posterior mean times for the most ancient nodes tended to increase as the number of partitions increases (Figure 6.2C). For example, the divergence time of the Metazoa vary from 833–681 Ma (single partition) to 834–787 Ma (10-partition; Table 6.1). Indeed, the closer to the root, the higher the discrepancy is, regardless of whether or not the nodes are calibrated (Figure 6.2C). Age estimates on intermediate nodes (e.g. all vertebrates and most arthropod nodes) do not vary significantly with partition strategy and for a small number of nodes, younger date estimates were obtained with increasing the number of partitions (Figure 6.2C, Table 6.1). Overall, nodes with highly variable time estimates among different partitions are those without calibration or are close to the root where the calibrations are less informative (Table D.4, Figure D.3).

Figure 6.4 shows the so-called infinite-sites plot in which the widths of the 95% HPD intervals are plotted against the posterior mean times (Rannala and Yang 2007). The scatter plot for the time prior shows high levels of uncertainty owing to the uncertain fossil calibrations: every 100My of divergence add 30My of uncertainty to the 95% prior interval width. The addition of molecular data increases precision substantially, and every 100My of divergence adds 18My of uncertainty to the posterior HPD interval (Figure 6.4B). The precision of node age estimates increases with the number of partitions. Dividing the data into more partitions gives narrower HPD widths, as indicated by the reduced regression coefficients in the plot. The extent of the reduction diminishes with higher numbers of partitions (for example, compare 4, 5 and 10 partitions), indicating that, given a fixed set of calibrations and fixed sequence data, the number of partitions is already near optimal in terms of dating precision. Nodes with the widest HPD intervals are those with no fossil

calibrations indicating that including more calibration points is likely to improve the precision of the time estimates. Finally, since the plots are very scattered (very low R^2 values), adding more sequence data may lead to more precise node age estimates.

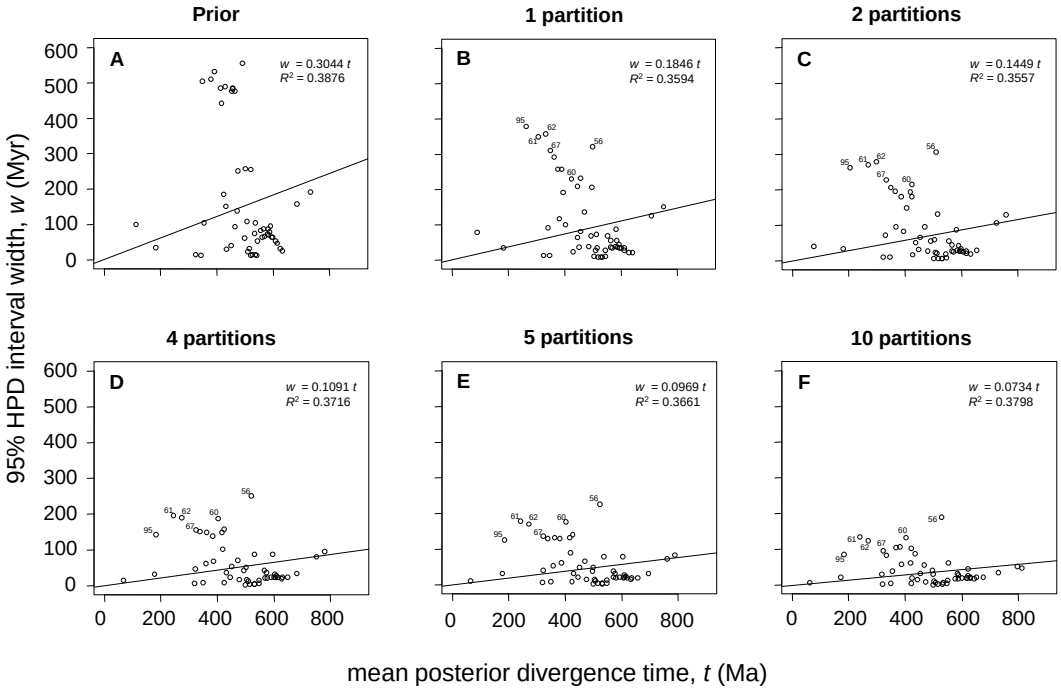

Figure 6.4: Infinite-sites plots. The 95% HPD width is plotted against the mean of the divergence times estimated without molecular data (prior) and with the 203 gene alignments divided into 1, 2, 4, 5, and 10 partitions under calibration strategy 1. The low correlations (R^2) indicate that the limited amount of sequence data contributes substantially to posterior uncertainty and the regression coefficients also indicate that the fossil calibrations involve much uncertainty. Node numbers are shown for nodes with the most uncertain time estimates.

6.3.4 The impact of phylogenetic uncertainty

In all previous analyses, a single tree topology has been used (Figure 6.6). However, the phylogenetic position of some metazoan taxa remains a subject of debate (Dunn, et al. 2014). To account for this uncertainty, we analysed 161 alternative binary trees accounting for uncertainties in the positioning of bilateria, chaetognaths, molluscs, nematodes, and xenacoelomorphs. The results show that nodes are affected differently depending on the tree topology. For example, some nodes have highly stable estimates across all topologies

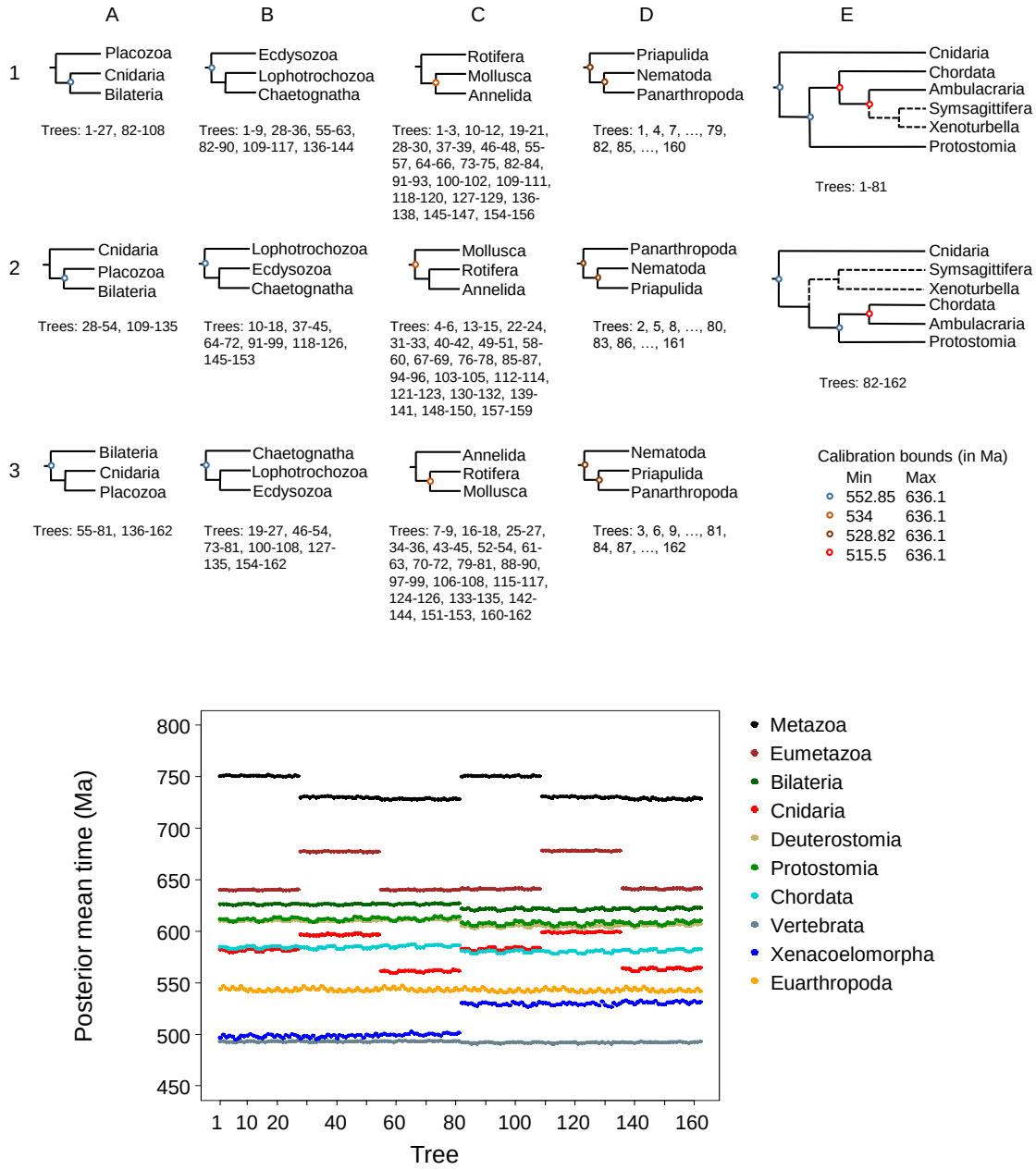

Figure 6.5: Effect of uncertainty in tree topology on divergence time estimates of the Metazoa. Four nodes (A-D) can be rearranged in three different ways (1-3) and a fifth node (E) can be rearranged in two ways resulting in a total of 162 tree topologies reflecting the uncertain relationships around these five nodes. Divergence times were estimated using calibration strategy 1, independent-rates model and 1 partition using each tree (bottom panel). Tree 1 is the main tree used in all other analyses. Some phylogenetic hypotheses had a strong effect on posterior mean times, for example, placing the Placozoa as the most basal with respect to Cnidaria and Bilateria (A), leads to substantially older divergence times for the Metazoa (bottom panel), while placing Cnidaria as the most basal leads to substantially older times for the divergence of Eumetazoa.

(Figure 6.5). These nodes are usually well calibrated and/or the local phylogeny well accepted, such as in deuterostomes and arthropods (Figure 6.5). In contrast, nodes with uncertain phylogenetic relationships exhibit considerable variation in estimated ages. These include the nodes close to the root of the tree such as Metazoa, Bilateria and Cnidaria; this variation increases with proximity to the root. For example, moving the position of Placozoa around the eumetazoan node has a profound impact on the estimated age of the root (Figure 6.5).

Other parameter settings such as the amino acid substitution model or the parameters of the birth-death model for the construction of time priors were found not to have a significant impact in time estimation.

6.4 Conclusions

There always has been a great interest in estimating the timeline of the evolution of life on earth and in particular that of animals. Powerful Bayesian methods based on the molecular clock methodology integrate information from molecular data and the fossil record, and have been promising in dealing with the issue. However, the results presented here indicate that the estimation of divergence times of the very early, ancient animal life forms via Bayesian molecular dating is extremely hard.

Our analysis integrated different interpretations of the animal fossil record in informing the minimum age of animal clades. Some of these identify fossil evidence of animals extending into the Cryogenian (Love, et al. 2009; Maloof, Rose, et al. 2010) while, at the other extreme, others argue that coherent evidence of animals is limited to the Cambrian, or the last few millions years of the Neoproterozoic (Budd and Jensen 2000). This is actually the long-standing conundrum of the Cambrian: whether the first animal fossils faithfully reflect an explosion in animal biodiversity, or merely an explosion of fossils (Runnegar 1982). In addition to discrepancies in interpretation of the early fossil record, the serious violation of the clock, the limited molecular data and uncertainties in the phylogenetic relationships of several metazoan taxa preclude the inference of a precise timeline of the metazoan evolution.

Composite results from our analyses, integrating major sources of uncertainties, indicate unequivocally that crown Metazoa originated 833–650 Ma in the Cryogenian, Eumetazoa 746–626 Ma, Bilateria 688–596 Ma, Deuterostomia 662–587 Ma, and Protostomia 653–578 Ma (Figure 6.6, Table 6.1). All last four groups diverged either in the late Cryogenian or the

early- to mid-Ediacaran. Those results suggest that the Cambrian Explosion is a phenomenon of fossilization, while biological diversity was established in the Neoproterozoic.

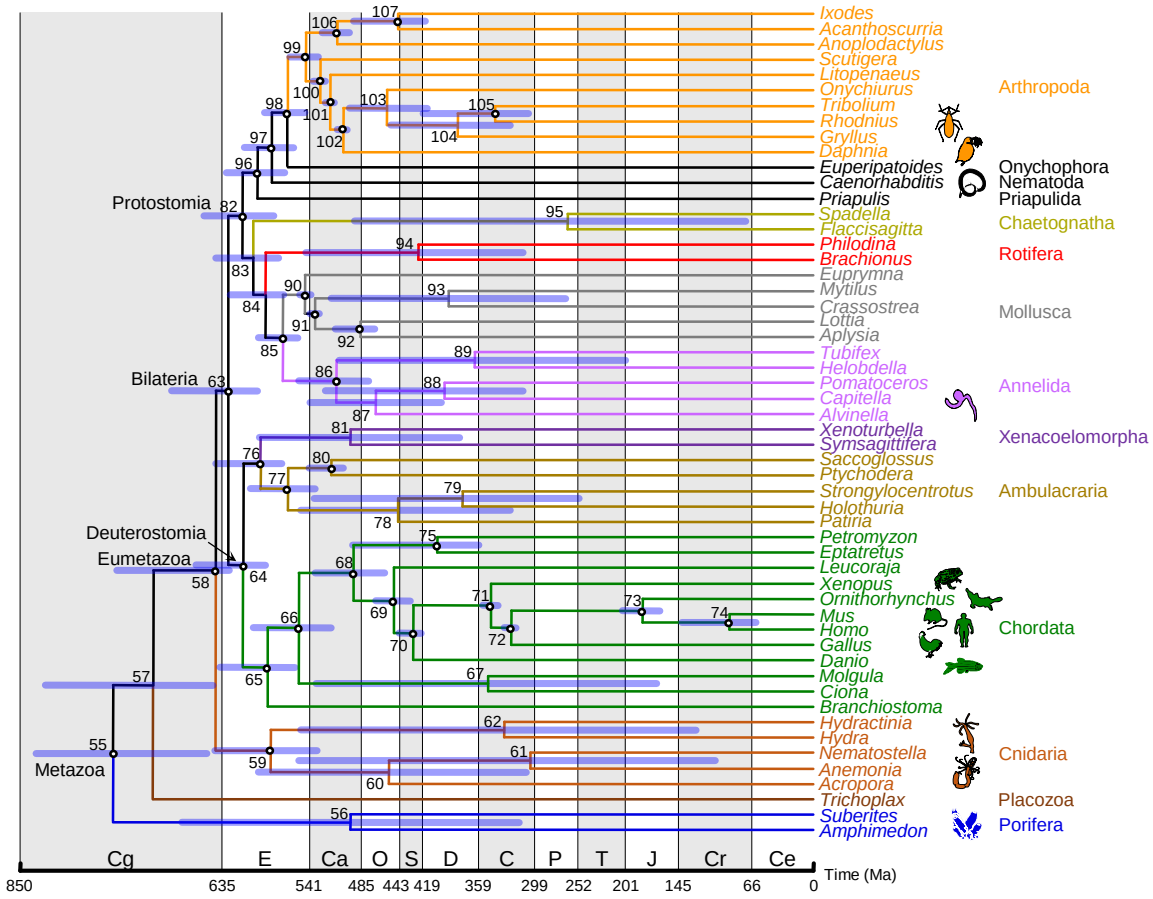

Figure 6.6: The timetree of the Metazoa encompassing major sources of uncertainty in time estimates. Node ages are plotted at the posterior mean for the analysis using calibration strategy 1, 1 partition, independent-rates model and LG+Γ_4+F substitution model. The node bars are composites extending from the minimum 2.5% HPD limit to the maximum 97.5% limit across all analyses (excluding results from calibration strategies 3 and 4 and from alternative topologies). Ce: Cenozoic, Cr: Cretaceous, J: Jurassic, T: Triassic, P: Permian, C: Carboniferous, D: Devonian, S: Silurian, O: Ordovician, Ca: Cambrian, E: Ediacaran, Cg: Cryogenian.

Our analysis allows, unequivocally, the rejection of the hypothesis that metazoans, eumetazoans, bilaterians, protostomes, deuterostomes, ecdysozoans, lophotrochozoans, or, for that matter, any of the major animal phyla, originated in the Cambrian. However, the uncertainties from competing interpretations of the fossil record, through the choice of rate models and sequence partition strategies, to competing phylogenetic hypotheses, all contribute to an evolutionary timescale that lacks sufficient precision to answer with confidence any other interesting hypotheses such as the relative divergence of protostomes

and deuterostomes. Some of this uncertainty can be reduced, for example, by adding more sequence data. Addition of molecular data will also help resolve phylogenetic relationships among specific metazoan taxa. However, the improvements in precision possible even with genome scale sequence data will be limited by the confounding effects of time and rate, which is the crux of the problem.

Thus attempts to build evolutionary narratives of animal evolution based on recent molecular clock studies appear to be premature. They fail to integrate different sources of uncertainties, which make accurate and precise divergence time estimates impossible with current data and methods. Nevertheless, some future progress might be possible through analysis of combined morphological and molecular data ("total-evidence" analysis) which uses morphological data of extant species to infer the placement of fossils on the phylogenetic tree, calibrating the tree at the same time. This combined analysis has been found to provide time estimates more precise and more robust to prior assumptions than traditional analyses based on fossil-based constraints (Ronquist, Klopfstein, et al. 2012). However, the morphology model used is very simplistic and improved models of morphological traits are required. We note that most combined analyses conducted to date have yielded unacceptably old divergence time estimates, even older than traditional node-calibrated studies (Ronquist, Klopfstein, et al. 2012; Arcila, et al. 2015). Thus much work remains to be done to elucidate the timeline of animal evolution on earth.

Summary

This thesis presented Bayesian methods to model sequence evolution and address important biological questions with particular emphasis on methods to study natural selection and species divergence times.

In Chapter 1 we provided an overview of the Bayesian theory and highlighted some key aspects necessary for the better understanding of the applications presented in the subsequent chapters.

In chapter 2 we presented some methods and popular programs for Bayesian phylogenetic analysis. Particular emphasis was given to Bayesian techniques to detect positive selection and estimate species divergence times integrating information from molecular data and the fossil record.

In Chapter 3 we developed a novel Bayesian method to estimate the nonsynonymous/ synonymous rate ratio and the sequence distance for pairwise comparisons of protein-coding gene sequences. Existing counting methods and the ML method based on a codon model of sequence evolution do not have nice statistical properties as they may return 0 or ∞ estimates in some data sets. In large genome-scale comparisons of protein-coding genes such extreme estimates are common and might cause difficulties in the calculation of summary statistics (e.g. mean, variance across all genes). In particular, the infinite estimates of ω are confusing to many users of the methods and ad hoc treatments are used to deal with the issue. The new Bayesian method always returns finite and reasonable estimates because the prior shrinks the posterior estimates away from extreme (0 or ∞) values and thus may provide a better procedure than ad hoc treatments which may introduce bias. Computer simulation and real data analyses revealed nice statistical properties for the Bayesian estimates (e.g. well defined, low MSEs). The Bayesian estimates are close to the MLEs when the data are informative, that is, when the sequences are long and the sequence divergence is intermediate. However, they can be quite different from the MLEs when the sequences are short and are either too similar (have little information about ω) or extremely divergent (contain too much noise about ω) because in those cases the prior has higher impact in posterior estimates. With informative data the power of the Bayesian method to detect positive selection (indicated by $\omega > 1$) is similar to that of the ML, but could be lower in case of uninformative data because of the prior. The effect of the prior decreases as the sequence length increases. The Bayesian method has been implemented in the CODEML program in the PAML package and is fast enough for genome scale comparisons of protein-coding gene sequences: a pair of sequences is analyzed in 1 to 2 seconds.

In chapter 4 we studied the impact of ancestral population size and incomplete lineage sorting in species divergence times estimated under the molecular clock with a Bayesian method which ignores the coalescent process. The coalescent process has long been recognised as an important aspect of molecular evolution however, the vast majority of molecular clock dating studies have ignored the effects of ancestral polymorphism and incomplete lineage sorting, probably because of the misconception that these aspects of evolution are only relevant to closely related species. We performed a combination of mathematical analysis, computer simulation and analysis of real data and we found that the estimates of divergence times and rate could be significantly biased when ancestral populations are large and when there is substantial incomplete lineage sorting. Divergence times are either over- or underestimated depending on the relative precision and configuration of fossil calibrations on the tree. For example, if the most informative calibrations are placed on the younger nodes of the phylogeny the ages of the internal nodes are underestimated, whereas if they are placed on the most ancient nodes the ages of the younger nodes are overestimated. In both cases the molecular rate is overestimated. We found that this is the case in both shallow and deep phylogenies with errors in deep phylogenies to be smaller. Although several Bayesian phylogenetic methods perform inference under the multi-species coalescent they are either computationally expensive or have been designed to work for only closely related species and are thus inappropriate to analyze the large genomic data currently available for many species. Further improvement of those methods could be advantageous to study species divergences.

In chapter 5 we evaluated the performance of five commonly used data partitioning strategies for the Bayesian estimation of species divergence times. In large genomic data sets it is important to account for variation in the evolutionary patterns across sites and partitioning is a commonly used approach. The method involves the grouping of sites that have been evolved under similar processes and the estimation of independent substitution models for each group. There are several ways to partition a data set into groups and the choice of partitioning scheme might affect the inference of divergence times. We used computer simulation and real data analysis to study differences in divergence time estimates using five partitioning strategies. In general, time estimates are similar among partitioning schemes, especially when the clock is not seriously violated, and thus not many safe conclusions can be made. The use of highly partitioned schemes reduces uncertainty of posterior estimates but accuracy may be poor when an incorrect rate-drift model is used. Automated tools to select the best-fit partitioning schemes, such as PartitionFinder, seem not to provide any advantage. Differences among partitioning schemes are larger when the clock is seriously violated. In an analysis of 78 plastid genes from 15 plant species where serious clock violation was detected, the time estimates varied substantially among partitioning

schemes, irrespective of the clock-model used. The similar performance of the partitioning schemes in most cases that we explored and some unexpected results with no obvious explanation precluded any further important conclusions, indicating that further research is needed.

In chapter 6 we applied a Bayesian algorithm implemented in the MCMCTREE program to estimate the timeline of animal evolution. Current molecular studies place the origin of Metazoa during the Cryogenian period which is further supported from recent fossil findings. Despite the consistency on the Cryogenian origin of the crown Metazoa, the evidence for the diversification of Bilateria remains controversial. Molecular studies place the origin of Bilateria during the Ediacaran period but the fossil record suggests a massive radiation of Bilateria after the Ediacaran-Cambrian boundary. Previous molecular dating studies have ignored the cumulative impact of several sources of uncertainty such as subjective interpretations of the fossil record, serious violation of the molecular clock, limited amount of molecular data and unresolved phylogenetic relationships among several taxa and have led to unduly precise time estimates. Our Bayesian dating analysis revealed that metazoan divergence time estimates are highly variable, largely depending on a series of analysis settings such as fossil calibrations, model of among-branches rate variation, data partitioning and tree topology. The analysis was based on 203 amino acid nuclear genes from 54 metazoan species in combination with 34 fossil calibrations. Although some of the uncertainty of time estimates can be reduced, for example by adding molecular data, this will be limited due to the confounding effect of rate and time possibly preventing a precise estimation of the animal evolutionary timescale. Recent alternative techniques based on a combined analysis of molecular and morphological data seem to be advantageous but more realistic models need to be developed for the accurate and precise estimation of the timescale of animal evolution.

Bayesian inference has become the basis of many phylogenetic results over the last years as many popular phylogenetic programs (e.g. MrBayes, BEAST) implement powerful MCMC algorithms and allow inference under complex and more realistic models. With this thesis I contribute in the development and proper use of Bayesian methods in molecular evolution. I developed a new Bayesian method to study natural selection, I examined the performance of a Bayesian method in estimating species divergence times and I used a Bayesian algorithm to estimate the divergence times of Metazoa highlighting important aspects for any study of species divergence times. I am optimistic that the proper use of existing Bayesian algorithms and the development of new more sophisticated Bayesian methods will help to shed light on important and interesting biological problems such as those examined here.

Appendices

A. Gaussian quadrature

Gaussian quadrature is a numerical integration method which uses Legendre polynomials to approximate any continuous integrand function $f(x)$. Because any polynomial is integrable analytically the integral is approximated as

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i), \tag{A.1}$$

where the weights w_i and the points x_i are predetermined given the total number of points n (Abramowitz and Stegun 1972, p. 887). The points are the roots of the n order Legendre polynomial $P_n(x)$ used to approximate the integrand while the weights are given by $w_i = \frac{2}{\left(1-x_i^2\right)\left[P_n'(x_i)\right]^2}$. Gaussian quadrature is optimal because if the integrand is a polynomial of degree $2n-1$ or less the method is exact. The number of points could be critical for the accuracy of the estimation. If the integrand is nearly flat over the interval (−1, 1) a few points are enough to approximate the integral reliably. In the extreme case that the integrand is perfectly flat even one point is enough for accurate estimation. However, if $f(x)$ is spiky and a small number of points is used the approximation might be quite poor and only if $f(x)$ is a low order polynomial the integral will be accurately estimated. In general the more points are used the better the approximation is but more calculations are necessary. Ideally, more points should be used in regions in which the integrand changes rapidly.

Legendre polynomials are defined in the interval [−1, 1] and thus the limits of integration are −1, 1. If the integration interval is different, say (a, b), this can be converted into the Gaussian-Legendre interval [−1, 1]. For example, using the transformation $u = \frac{b-a}{2}x + \frac{a+b}{2}$, where $-1 < x < 1$ and $a < u < b$ and since $\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b-a}{2}$ we have

$$\int_{a}^{b} f(u)\mathrm{d}u = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}x + \frac{a+b}{2}\right)dx \approx \frac{b-a}{2}\sum_{i=1}^{n} w_i f\left(\frac{b-a}{2}x_i + \frac{a+b}{2}\right). \tag{A.2}$$

Some mappings (a, b) to (−1, 1) might be more efficient than others as they might give a transformed integrand more flat in the interval (−1, 1) and thus the same accuracy would be achieved with fewer points. A proper transformation can then offer computational advantage.

Similarly we can estimate a two-dimensional integral as

$$\int_a^b \int_c^d f(x,y)\mathrm{d}x\mathrm{d}y \approx \sum_{i,j=1}^{n} w_i w_j r(u_i,\upsilon_j), \tag{A.3}$$

where $r(u,\upsilon) = f(x,y)|J|$ with $J = \left|\frac{\partial(x,y)}{\partial(u,\upsilon)}\right|$ to be the Jacobian determinant of the transform $x = x(u,\upsilon)$, $y = y(u,\upsilon)$ and u_i, u_j, w_i, w_j to be pre-determined given the total number of points n in each dimension. Different numbers of points may be used in the two dimensions. Note that for d-dimensional integrals the computation is proportional to n^d, making the calculation of high dimensional integrals ($d > 3$) practically impossible.

B. Estimating the variance of ω and t using the Nei & Gojobori method

According to Nei and Gojobori's (1986) method given an alignment of two sequences the numbers of synonymous (d_S) and nonsynonymous (d_N) differences per site are given by

$$\begin{aligned} \hat{d}_S &= -\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}_S\right), \\ \hat{d}_N &= -\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}_N\right), \end{aligned} \tag{B.1}$$

where $\hat{p}_S = \frac{S_d}{S}$ and $\hat{p}_N = \frac{N_d}{N}$ are the observed proportions of synonymous and nonsynonymous differences per site. The S_d, N_d, S, N are calculated as Nei and Gojobori (1986) proposed (see §3.1). We assume that $\hat{p}_S$ and $\hat{p}_N$ are independent binomial proportions and thus

$$\begin{aligned} \hat{V}(\hat{p}_S) &= \frac{\hat{p}_S(1-\hat{p}_S)}{S}, \\ \hat{V}(\hat{p}_N) &= \frac{\hat{p}_N(1-\hat{p}_N)}{N}, \\ Cov(\hat{p}_S,\hat{p}_N) &= 0. \end{aligned} \tag{B.2}$$

We use the Delta method to estimate the variances $V\left(\hat{d}_S\right)$ and $V\left(\hat{d}_N\right)$. The Delta method states that given a random variable x with mean μ_x and variance σ_x^2, the variance of the random variable $y = g(x)$ is given by $V(y) \approx \sigma_x^2\left[\frac{dg(\mu_x)}{dx}\right]^2$. Thus

$$\hat{V}\left(\hat{d}_S\right)=\hat{V}\left(-\frac{3}{4}\log\left(1-\frac{4}{3}\hat{p}_S\right)\right)\approx\left[\frac{1}{1-\frac{4}{3}\hat{p}_S}\right]^2\hat{V}\left(\hat{p}_S\right), \tag{B.3}$$

and

$$\hat{V}\left(\hat{d}_N\right)\approx\left[\frac{1}{1-\frac{4}{3}\hat{p}_N}\right]^2\hat{V}\left(\hat{p}_N\right), \tag{B.4}$$

where $\hat{V}\left(\hat{p}_S\right)$ and $\hat{V}\left(\hat{p}_N\right)$ are given by (B.2).

Given two random variables *x*, *y* with means μ_x, μ_y and variances σ_x^2, σ_y^2, the variance of the random variable $z = x/y$, according to the Delta technique, is $Var\left(\frac{x}{y}\right)\approx\frac{\sigma_x^2}{\mu_y^2}-\frac{2\mu_x\sigma_{xy}}{\mu_y^3}+\frac{\mu_x^2\sigma_y^2}{\mu_y^4}$ where σ_{xy} is the covariance of *x* and *y*. Thus, since $\omega=\frac{d_N}{d_S}$, we have

$$\hat{V}\left(\hat{\omega}\right)=\hat{V}\left(\frac{\hat{d}_N}{\hat{d}_S}\right)\approx\frac{\hat{V}\left(\hat{d}_N\right)}{\hat{d}_S^2}-\frac{2\hat{d}_N Cov\left(\hat{d}_N,\hat{d}_S\right)}{\hat{d}_S^3}+\frac{\hat{d}_N^2\hat{V}\left(\hat{d}_S\right)}{\hat{d}_S^4}=...$$

$$...=\left[\frac{4\log\left(1-\frac{4}{3}\hat{p}_N\right)}{3\left(1-\frac{4}{3}\hat{p}_S\right)\left[\log\left(1-\frac{4}{3}\hat{p}_S\right)\right]^2}\right]^2\hat{V}\left(\hat{p}_S\right)+\left[\frac{4}{3\left(1-\frac{4}{3}\hat{p}_N\right)\log\left(1-\frac{4}{3}\hat{p}_S\right)}\right]^2\hat{V}\left(\hat{p}_N\right), \tag{B.5}$$

with $\hat{V}\left(\hat{p}_S\right)$ and $\hat{V}\left(\hat{p}_N\right)$ as in (B.2).

The distance *t* according to the Nei and Gojobori (1986) method is given by

$$t=\frac{3S}{S+N}d_S+\frac{3N}{S+N}d_N. \tag{B.6}$$

Thus the estimate of the distance variance is

$$\hat{V}\left(\hat{t}\right)=\hat{V}\left(\frac{3S}{S+N}\hat{d}_S+\frac{3N}{S+N}\hat{d}_N\right)\approx\left[\frac{3S}{S+N}\frac{1}{1-\frac{4}{3}\hat{p}_S}\right]^2\hat{V}\left(\hat{p}_S\right)+\left[\frac{3N}{S+N}\frac{1}{1-\frac{4}{3}\hat{p}_N}\right]^2\hat{V}\left(\hat{p}_N\right) \tag{B.7}$$

C. Calculating $P(\omega > 1 \mid x)$ using Gaussian quadrature

We are interested in calculating the posterior probability for $\omega > 1$ for a pairwise sequence alignment. This is given by

$$P(\omega > 1 \mid x) = \frac{1}{C}\int_0^\infty\int_1^\infty f(x \mid t, \omega, \hat{\kappa})\, f(t, \omega)\mathrm{d}t\mathrm{d}\omega, \tag{C.1}$$

where $f(x|t,\omega,\hat{\kappa})$ is the likelihood, $f(x|t,\omega)$ is the joint prior on ω and t and, C is the normalizing constant. Following similar techniques to §3.2 the integral $I = \int_0^\infty\int_1^\infty f(x \mid t, \omega, \hat{\kappa})\, f(t, \omega)\mathrm{d}t\mathrm{d}\omega$ becomes $I = e^{l_{\max}}\int_0^\infty\int_1^\infty h(t, \omega)\mathrm{d}t\mathrm{d}\omega$, where $h(t, \omega)$ and $l_{\max}$ are as in §3.2.

We assume that $x_1 = \log t \sim \text{Logistic}(\mu_1, \sigma_1)$ and $x_2 = \log\omega \sim \text{Logistic}(\mu_2, \sigma_2)$ and we perform the following change of variables:

$$z_1 = 2F_L(x_1) - 1 \Rightarrow t = \exp\left\{\mu_1 + \sigma_1 \log\frac{1+z_1}{1-z_1}\right\}, \tag{C.2}$$

$$u = 2F_L(x_2) - 1 \Rightarrow \omega = \exp\left\{\mu_2 + \sigma_2 \log\frac{1+u}{1-u}\right\}, \tag{C.3}$$

where $F_L(x)$ is the CDF of the logistic distribution. Thus the integral becomes

$I = exp(l_{max})\int_{-1}^{1}\int_{a}^{1} h(t,\omega)\frac{2t\sigma_1}{1-z_1^2}\frac{2\omega\sigma_2}{1-u^2}\mathrm{d}u\mathrm{d}z_1$, where $a = 2F_L(0) - 1$. We then perform the following transform to map the interval $(a, 1)$ to $(-1, 1)$:

$$u = \frac{1-a}{2}z_2 + \frac{1+a}{2} \Rightarrow z_2 = \frac{2}{1-a}u - \frac{1+a}{1-a}. \tag{C.4}$$

Thus the integral becomes

$I = exp(l_{max})\int_{-1}^{1}\int_{-1}^{1} h(t,\omega)\frac{2t\sigma_1}{1-z_1^2}\frac{2\omega\sigma_2}{1-u^2}\frac{1-a}{2}\mathrm{d}z_1\mathrm{d}z_2 \approx exp(l_{max})\sum_{i,j} w_i w_j q(z_{1_i}, z_{2_j})$, where t, ω, u are given by (C.2), (C.3), (C.4), respectively. Then $P(\omega > 1 \mid x) \approx \frac{1}{A}\sum_{i,j=1}^{n} w_i w_j r(z_{1_i}, z_{2_j})$, where $A = C\exp(-l_{\max})$ and C is as in (3.13). The constant term $\exp(l_{\max})$ cancels during calculations.

D. Supplementary tables and figures for chapter 6

Here are additional tables and figures concerning the estimation of divergence times of Metazoa described in chapter 6.

Table D.1: Fossil calibration densities constructed from the minimum and maximum constrains.

Node	Clade	Min	Max	Strategy 1	Strategy 2	Strategy 3	Strategy 4
55	Metazoa	552.85	833	B(5.5285,8.33,0.001,0.001)	B(5.5285,8.33,0.001,0.001)	***B(6.349,8.33,0.001,0.001)***	***B(6.349,8.33,0.001,0.001)***
58	Eumetazoa	552.85	636.1	B(5.5285,6.361,0.001,0.025)	***SN(5.6,0.34,7)***	***L(5.5285,0,10,0.001)***	***L(5.5285,0,0.1,0.001)***
59	Cnidaria	529	636.1	B(5.29,6.361,0.001,0.025)	***SN(5.38,0.44,7)***	***L(5.29,0,10,0.001)***	***L(5.29,0,0.1,0.001)***
63	Bilateria	552.85	636.1	B(5.5285,6.361,0.001,0.025)	***SN(5.6,0.34,7)***	***L(5.5285,0,10,0.001)***	***L(5.5285,0,0.1,0.001)***
64	Deuterostomia	515.5	636.1	B(5.155,6.361,0.001,0.025)	***SN(5.255,0.5,7)***	***L(5.155,0,10,0.001)***	***L(5.155,0,0.1,0.001)***
65	Chordata	514	636.1	B(5.14,6.361,0.001,0.025)	***SN(5.25,0.5,7)***	***L(5.14,0,10,0.001)***	***L(5.14,0,0.1,0.001)***
66	Olfactores	514	636.1	B(5.14,6.361,0.001,0.025)	***SN(5.25,0.5,7)***	***L(5.14,0,10,0.001)***	***L(5.14,0,0.1,0.001)***
68	Vertebrata	457.5	636.1	B(4.575,6.361,0.001,0.025)	***SN(4.7,0.75,9)***	***L(4.575,0,10,0.001)***	***L(4.575,0,0.1,0.001)***
69	Gnathostomata	420.7	468.4	B(4.207,4.684,0.001,0.025)	B(4.207,4.684,0.001,0.025)	B(4.207,4.684,0.001,0.025)	B(4.207,4.684,0.001,0.025)
70	Osteichthyes	420.7	453.7	B(4.207,4.537,0.001,0.025)	B(4.207,4.537,0.001,0.025)	B(4.207,4.537,0.001,0.025)	B(4.207,4.537,0.001,0.025)
71	Tetrapoda	337	351	B(3.37,3.51,0.001,0.025)	B(3.37,3.51,0.001,0.025)	B(3.37,3.51,0.001,0.025)	B(3.37,3.51,0.001,0.025)
72	Amniota	318	332.9	B(3.18,3.329,0.001,0.025)	B(3.18,3.329,0.001,0.025)	B(3.18,3.329,0.001,0.025)	B(3.18,3.329,0.001,0.025)
73	Mammalia	164.9	201.5	B(1.649,2.015,0.001,0.025)	B(1.649,2.015,0.001,0.025)	B(1.649,2.015,0.001,0.025)	B(1.649,2.015,0.001,0.025)
74	Euarchontoglires	61.6	164.6	B(0.616,1.646,0.001,0.025)	B(0.616,1.646,0.001,0.025)	B(0.616,1.646,0.001,0.025)	B(0.616,1.646,0.001,0.025)
75	Cyclostomata	358.5	636.1	B(3.585,6.361,0.001,0.025)	B(3.585,6.361,0.001,0.025)	B(3.585,6.361,0.001,0.025)	B(3.585,6.361,0.001,0.025)
76	Xenambulacraria	515.5	636.1	B(5.155,6.361,0.001,0.025)	B(5.155,6.361,0.001,0.025)	B(5.155,6.361,0.001,0.025)	B(5.155,6.361,0.001,0.025)
77	Ambulacraria	515.5	636.1	B(5.155,6.361,0.001,0.025)	B(5.155,6.361,0.001,0.025)	B(5.155,6.361,0.001,0.025)	B(5.155,6.361,0.001,0.025)
80	Hemichordata	504.5	636.1	B(5.045,6.361,0.001,0.025)	B(5.045,6.361,0.001,0.025)	B(5.045,6.361,0.001,0.025)	B(5.045,6.361,0.001,0.025)
82	Protostomia	552.85	636.1	B(5.5285,6.361,0.001,0.025)	***SN(5.6,0.34,7)***	***L(5.5285,0,10,0.001)***	***L(5.5285,0,0.1,0.001)***
85	Annelids-Molluscs	534	636.1	B(5.34,6.361,0.001,0.025)	***SN(5.41,0.43,9)***	***L(5.34,0,10,0.001)***	***L(5.34,0,0.1,0.001)***
86	Capitellid-Polychete-leech	476.5	636.1	B(4.765,6.361,0.001,0.025)	***SN(4.86,0.68,10)***	***L(4.765,0,10,0.001)***	***L(4.765,0,0.1,0.001)***
90	Mollusca	534	549	B(5.34,5.49,0.001,0.025)	B(5.34,5.49,0.001,0.025)	B(5.34,5.49,0.001,0.025)	B(5.34,5.49,0.001,0.025)
91	Bivalve-Gastropod	530	549	B(5.30,5.49,0.001,0.025)	B(5.30,5.49,0.001,0.025)	B(5.30,5.49,0.001,0.025)	B(5.30,5.49,0.001,0.025)
92	Gastropoda	470.2	549	B(4.702,5.49,0.001,0.025)	***SN(4.75,0.33,9)***	***L(4.702,0,10,0.001)***	***L(4.702,0,0.1,0.001)***
96	Ecdysozoa	528.82	636.1	B(5.2882,6.361,0.001,0.025)	B(5.2882,6.361,0.001,0.025)	B(5.2882,6.361,0.001,0.025)	B(5.2882,6.361,0.001,0.025)
97	Nematoda-Arthropoda	528.82	636.1	B(5.2882,6.361,0.001,0.025)	***SN(5.38,0.44,7)***	***L(5.2882,0,10,0.001)***	***L(5.2882,0,0.1,0.001)***
98	Lobopodia	528.82	636.1	B(5.2882,6.361,0.001,0.025)	B(5.2882,6.361,0.001,0.025)	B(5.2882,6.361,0.001,0.025)	B(5.2882,6.361,0.001,0.025)
99	Euarthropoda	514	636.1	B(5.14,6.361,0.001,0.025)	***SN(5.22,0.52,9)***	***L(5.14,0,10,0.001)***	***L(5.14,0,0.1,0.001)***
100	Mandibulata	514	531.22	B(5.14,5.3122,0.001,0.025)	B(5.14,5.3122,0.001,0.025)	B(5.14,5.3122,0.001,0.025)	B(5.14,5.3122,0.001,0.025)
101	Pancrustacea	514	531.22	B(5.14,5.3122,0.001,0.025)	B(5.14,5.3122,0.001,0.025)	B(5.14,5.3122,0.001,0.025)	B(5.14,5.3122,0.001,0.025)
102	Copepoda-Branchiopoda	499	531.22	B(4.99,5.3122,0.001,0.025)	B(4.99,5.3122,0.001,0.025)	B(4.99,5.3122,0.001,0.025)	B(4.99,5.3122,0.001,0.025)
105	Eumetabola	305.5	413.6	B(3.055,4.136,0.001,0.025)	B(3.055,4.136,0.001,0.025)	B(3.055,4.136,0.001,0.025)	B(3.055,4.136,0.001,0.025)
106	Pycnogonida-other chelicertates	497.5	531.22	B(4.975,5.3122,0.001,0.025)	B(4.975,5.3122,0.001,0.025)	B(4.975,5.3122,0.001,0.025)	B(4.975,5.3122,0.001,0.025)
107	Acari-Arenacea	416	531.22	B(4.16,5.3122,0.001,0.025)	B(4.16,5.3122,0.001,0.025)	B(4.16,5.3122,0.001,0.025)	B(4.16,5.3122,0.001,0.025)

Note.− B(t_L, t_U, p_L, p_U) means the node age has a soft uniform distribution between a minimum time t_L and a maximum time t_U, with probabilities p_L and p_U that the age is outside the bounds. SN(t, a, b) means the node age has a skew-normal distribution with location t, scale a, and shape b. L(t_L, p, c, p_L) means that the node age has a Cauchy distribution truncated on the left at t_L, with mode parameter p, tail parameter c, and probability p_L that the node age is younger than the minimum bound. Nodes and calibration densities that are different among the calibration strategies are indicated with bold typeface and italics.

Table D.2: Minimum and maximum fossil constraints and 95% interval of prior divergence times (Ma) for all metazoan clades under the four calibration strategies.

Node	Crown group	Min	Max	S1, IR, 1P		S2, IR, 1P		S3, IR, 1P		S4, IR, 1P	
55	Metazoa	552.85	833	641.3	832.6	629.5	833.2	757.5	833.5	689.0	833.3
56				123.1	679.0	115.6	664.5	160.0	826.9	129.8	768.3
57				622.4	778.9	599.9	777.8	738.2	832.5	662.6	826.7
58	Eumetazoa	552.85	636.1	616.6	642.7	589.5	658.7	717.4	830.5	629.4	806.7
59	Cnidaria	529	636.1	538.8	634.8	536.7	615.8	543.5	795.1	529.0	687.5
60				371.8	630.3	373.7	609.4	332.1	702.7	336.6	625.1
61				49.2	577.5	58.7	567.3	53.8	616.1	50.1	572.1
62				139.9	626.1	135.2	602.5	108.4	704.4	121.3	631.0
63	Bilateria	552.85	636.1	605.4	637.6	579.6	635.6	677.0	817.9	598.3	751.5
64	Deuterostomia	515.5	636.1	581.0	633.7	558.2	618.9	618.4	785.4	564.5	695.3
65	Chordata	514	636.1	535.2	623.1	532.1	596.3	546.0	749.5	519.0	630.1
66	Olfactores	514	636.1	513.9	598.1	518.1	576.0	506.8	696.9	513.9	584.2
67				107.8	592.6	112.6	573.8	110.6	641.8	89.7	570.1
68	Vertebrata	457.5	636.1	457.4	565.4	464.6	545.7	451.9	633.4	457.4	532.4
69	Gnathostomata	420.7	468.4	429.6	469.6	430.1	470.1	429.6	469.6	429.8	469.5
70	Osteichthyes	420.7	453.7	420.7	451.2	420.7	451.4	420.7	451.2	420.6	451.2
71	Tetrapoda	337	351	337.2	350.9	337.2	350.9	337.2	350.9	337.4	351.0
72	Amniota	318	332.9	318.3	332.9	318.2	332.7	318.3	332.8	318.1	332.6
73	Mammalia	164.9	201.5	165.6	201.2	165.5	201.2	165.9	201.5	165.6	201.2
74	Euarchontoglires	61.6	164.6	63.2	163.6	63.9	163.8	63.5	163.6	63.3	163.5
75	Cyclostomata	358.5	636.1	358.1	509.9	358.3	500.6	358.1	539.8	358.2	491.4
76	Xenambulacraria	515.5	636.1	547.0	625.3	534.4	605.8	561.6	643.2	548.2	638.8
77	Ambulacraria	515.5	636.1	519.3	605.8	516.1	586.7	526.8	630.5	517.6	616.2
78				335.9	592.6	338.9	578.2	331.8	607.7	334.7	595.2
79				45.3	556.6	51.0	552.9	34.7	550.4	50.1	556.5
80	Hemichordata	504.5	636.1	504.2	577.4	504.2	563.8	504.3	593.7	504.3	584.4
81				133.2	613.3	136.6	595.9	140.3	634.7	146.0	628.0
82	Protostomia	552.85	636.1	587.2	634.4	567.3	619.5	620.8	786.0	573.6	693.7
83				563.3	628.8	552.0	608.3	572.3	756.5	551.3	655.7
84				548.9	619.4	545.1	597.7	549.2	726.2	543.4	631.5
85	Annelids-Molluscs	534	636.1	539.4	605.5	539.8	582.6	539.6	693.8	536.4	601.8
86	Capitellid-Polychete-Leech	476.5	636.1	476.7	581.5	487.2	566.3	471.7	642.7	476.5	565.5
87				321.4	570.8	323.4	557.4	321.0	607.6	315.4	559.4
88				39.9	542.9	41.3	538.2	37.3	545.9	37.1	534.8
89				85.5	567.6	90.9	557.9	70.6	591.6	82.0	558.7
90	Mollusca	534	549	535.2	549.3	535.2	549.3	535.4	549.5	535.0	549.2
91	Bivalve-Gastropod	530	549	530.0	545.2	530.0	544.7	530.0	545.3	530.0	544.9
92	Gastropoda	470.2	549	470.2	532.7	472.3	527.0	460.7	536.8	470.1	528.2
93				106.6	544.9	106.7	545.5	100.6	544.4	108.4	545.5
94				134.0	613.7	134.2	592.0	104.0	646.5	126.1	612.9
95				147.7	623.1	140.4	600.8	110.5	685.0	137.9	634.0
96	Ecdysozoa	528.82	636.1	562.8	627.6	551.4	607.6	575.8	641.6	560.3	638.2
97	Nematoda-Arthropoda	528.82	636.1	543.0	614.7	539.7	591.8	551.7	634.0	538.2	617.8
98	Lobopodia	528.82	636.1	529.7	595.0	529.3	577.1	531.7	613.4	528.8	594.2
99	Euarthropoda	514	636.1	520.8	574.9	522.3	560.6	521.3	588.1	519.5	567.8
100	Mandibulata	514	531.22	517.6	532.0	517.6	531.8	517.6	531.9	517.3	531.7
101	Pancrustacea	514	531.22	514.0	528.2	514.0	528.1	514.0	528.3	514.0	528.0
102	Copepoda-Branchiopoda	499	531.22	499.0	522.3	499.0	522.1	499.0	522.4	499.0	522.0
103				388.4	524.7	393.1	525.3	386.4	524.9	396.9	524.7
104				321.7	507.1	323.4	509.4	319.4	505.1	322.4	506.8
105	Eumetabola	305.5	413.6	305.4	409.2	305.4	409.7	305.5	409.6	305.4	409.1
106	Pcynogonida-other chelicertates	497.5	531.22	497.5	529.9	497.5	529.8	497.6	530.1	497.5	529.7
107	Acari-Arenacea	416	531.22	415.9	509.7	416.0	509.8	416.0	509.3	416.0	509.6

Note.− Prior times are 95% intervals estimated by running MCMCTREE without sequence data under the four calibration strategies S1–S4. IR: Independent-rates model. 1P: The 203 proteins analysed as a single partition. Node numbers are as in Figure 6.6.

Table D.3: Minimum and maximum fossil constraints and 95% HPD interval of posterior divergence times (Ma) for all metazoan clades under the four calibration strategies.

Node	Crown group	Min	Max	S1, IR, 1P		S2, IR, 1P		S3, IR, 1P		S4, IR, 1P	
55	Metazoa	552.85	833	680.6	832.7	716.2	833.4	795.2	833.6	780.0	833.5
56				314.6	639.9	318.6	646.6	319.2	670.6	319.2	661.1
57				649.2	776.7	686.1	805.5	779.5	832.2	761.6	831.4
58	Eumetazoa	552.85	636.1	630.7	652.9	649.5	714.2	738.5	808.8	715.4	798.7
59	Cnidaria	529	636.1	533.3	620.5	537.7	631.9	583.8	760.0	531.5	715.7
60				318.9	554.4	319.3	550.0	350.7	637.9	319.6	591.2
61				110.4	458.5	118.1	452.7	126.1	475.4	129.2	461.8
62				125.5	488.1	133.8	485.5	188.3	542.2	167.2	519.5
63	Bilateria	552.85	636.1	615.1	637.8	624.2	672.3	685.4	759.2	666.4	736.4
64	Deuterostomia	515.5	636.1	593.7	627.9	598.0	649.6	643.7	721.7	625.9	695.3
65	Chordata	514	636.1	555.4	611.3	558.1	622.2	600.5	693.3	568.6	662.6
66	Olfactores	514	636.1	516.6	583.6	524.3	588.0	548.2	656.1	521.8	618.6
67				167.9	480.9	193.2	485.3	236.2	526.3	203.9	486.4
68	Vertebrata	457.5	636.1	459.6	527.9	467.1	527.6	469.2	564.7	461.8	533.5
69	Gnathostomata	420.7	468.4	432.9	468.7	433.9	468.6	435.9	469.4	433.8	468.4
70	Osteichthyes	420.7	453.7	420.6	444.1	420.6	443.9	420.6	443.6	420.6	441.9
71	Tetrapoda	337	351	338.3	351.4	338.4	351.5	338.8	351.6	338.7	351.6
72	Amniota	318	332.9	318.0	331.4	318.0	331.1	318.0	330.7	318.0	330.7
73	Mammalia	164.9	201.5	165.1	200.7	164.9	200.5	164.9	200.6	165.0	200.5
74	Euarchontoglires	61.6	164.6	61.4	140.2	61.4	135.3	61.4	127.6	61.3	128.4
75	Cyclostomata	358.5	636.1	358.1	458.0	358.1	455.8	358.1	469.1	358.1	453.0
76	Xenambulacraria	515.5	636.1	569.8	614.5	575.9	632.2	606.4	646.4	600.6	644.4
77	Ambulacraria	515.5	636.1	534.6	591.3	538.5	603.5	554.8	620.1	552.7	618.8
78				330.6	537.8	334.3	541.3	348.9	550.0	343.3	548.1
79				250.6	507.0	266.4	509.1	285.6	510.8	277.5	508.1
80	Hemichordata	504.5	636.1	504.2	537.6	504.2	540.0	504.1	545.6	504.2	546.2
81				378.5	585.8	404.9	594.0	421.7	605.3	420.1	605.0
82	Protostomia	552.85	636.1	598.0	626.4	603.6	647.5	644.4	712.3	632.2	690.5
83				582.7	616.2	587.6	633.1	620.3	693.2	610.6	672.4
84				570.0	605.7	573.7	618.3	596.5	671.0	588.8	649.1
85	Annelids-Molluscs	534	636.1	552.3	586.1	554.1	591.7	564.2	630.1	559.5	611.9
86	Capitellid-Polychete-Leech	476.5	636.1	476.3	548.1	480.9	550.9	468.3	573.6	476.4	550.0
87				398.5	536.0	407.4	534.2	413.1	548.3	406.2	533.6
88				310.5	501.1	312.3	489.6	315.3	499.9	312.1	481.2
89				201.6	487.0	220.6	485.2	231.3	482.9	226.7	473.5
90	Mollusca	534	549	538.4	549.6	539.1	549.7	540.8	550.0	540.5	550.0
91	Bivalve-Gastropod	530	549	530.0	539.1	530.0	538.6	530.0	538.2	530.0	538.3
92	Gastropoda	470.2	549	470.0	508.3	470.3	506.2	450.8	505.3	470.1	500.9
93				265.0	516.5	285.1	512.1	300.3	505.4	291.9	507.6
94				310.6	541.4	314.2	538.2	319.4	549.4	318.5	544.6
95				72.7	452.4	84.3	447.5	93.3	454.7	88.8	447.7
96	Ecdysozoa	528.82	636.1	577.8	613.2	581.9	627.1	610.1	644.5	602.6	641.6
97	Nematoda-Arthropoda	528.82	636.1	561.4	599.8	563.8	608.3	583.9	628.6	577.8	625.0
98	Lobopodia	528.82	636.1	545.1	582.8	547.8	588.5	558.5	606.1	554.7	602.0
99	Euarthropoda	514	636.1	530.8	559.4	531.9	560.7	535.4	571.0	534.5	567.1
100	Mandibulata	514	531.22	523.4	532.3	524.0	532.3	525.2	532.6	525.0	532.4
101	Pancrustacea	514	531.22	514.0	522.8	514.0	522.3	514.0	521.8	514.0	521.9
102	Copepoda-Branchiopoda	499	531.22	499.0	510.1	498.9	509.2	498.9	508.0	498.9	508.3
103				414.4	496.1	414.2	493.6	418.0	490.3	417.6	491.5
104				324.8	441.5	325.3	438.8	327.0	433.4	327.6	435.5
105	Eumetabola	305.5	413.6	305.3	396.8	305.3	393.1	305.2	387.2	305.2	388.3
106	Pcynogonida-other chelicertates	497.5	531.22	497.5	526.1	497.5	525.8	497.5	526.9	497.5	526.4
107	Acari-Arenacea	416	531.22	415.9	479.9	415.8	477.5	415.9	474.0	415.8	474.4

Note.− Posterior times are the 95% HPD intervals estimated with MCMCTREE under the LG+Γ_4+F model, using four calibration strategies S1–S4. IR: Independent-rates model. 1P: The 203 proteins analysed as a single partition. Node numbers are as in Figure 6.6.

Table D.4: 95% HPD interval of posterior divergence times (Ma) for all metazoan clades under various partitioning schemes.

Node	Crown group	S1, IR, 1P		S1, IR, 2P		S1, IR, 4P		S1, IR, 5P		S4, IR, 10P	
55	Metazoa	680.6	832.7	701.0	831.2	736.9	832.6	748.9	832.3	786.8	833.5
56		314.6	639.9	326.2	632.6	387.9	639.0	413.3	639.2	440.3	631.0
57		649.2	776.7	674.7	781.6	712.5	794.3	726.8	798.5	771.7	823.1
58	Eumetazoa	630.7	652.9	638.9	669.0	664.2	699.9	677.5	711.0	712.2	746.2
59	Cnidaria	533.3	620.5	532.7	620.2	548.7	635.6	559.0	637.4	596.2	641.7
60		318.9	554.4	315.9	531.3	315.7	501.8	310.6	487.4	335.4	469.1
61		110.4	458.5	119.5	388.5	154.1	349.4	159.3	336.3	186.2	320.4
62		125.5	488.1	155.3	432.4	176.7	365.6	187.8	358.0	207.0	331.1
63	Bilateria	615.1	637.8	623.1	643.3	636.6	660.0	646.4	666.5	665.6	688.3
64	Deuterostomia	593.7	627.9	602.2	630.1	617.0	640.5	624.2	644.0	639.5	662.3
65	Chordata	555.4	611.3	567.9	611.3	586.2	619.3	593.4	621.6	609.0	635.7
66	Olfactores	516.6	583.6	527.9	584.5	544.0	586.5	552.0	589.6	568.0	600.0
67		167.9	480.9	215.4	444.5	233.7	391.3	246.4	383.9	274.6	371.0
68	Vertebrata	459.6	527.9	464.5	520.8	472.3	515.6	475.4	514.4	483.3	512.9
69	Gnathostomata	432.9	468.7	432.5	464.7	433.6	457.4	434.8	456.2	436.2	451.3
70	Osteichthyes	420.6	444.1	420.6	437.8	420.6	430.6	420.6	428.8	420.6	425.0
71	Tetrapoda	338.3	351.4	339.8	351.7	342.7	351.9	343.8	352.0	346.5	352.1
72	Amniota	318.0	331.4	318.0	329.3	318.0	325.2	318.0	323.9	318.0	321.5
73	Mammalia	165.1	200.7	164.9	200.0	164.8	197.8	164.8	196.4	164.8	186.5
74	Euarchontoglires	61.4	140.2	61.3	102.8	61.3	76.8	61.3	73.2	61.3	67.3
75	Cyclostomata	358.1	458.0	358.3	442.2	358.1	426.1	358.2	420.4	358.3	416.5
76	Xenambulacraria	569.8	614.5	580.7	615.2	595.1	623.7	601.7	626.6	617.6	639.9
77	Ambulacraria	534.6	591.3	542.9	588.2	555.1	591.4	559.6	592.3	572.6	600.1
78		330.6	537.8	331.9	516.0	341.2	488.9	348.9	481.5	367.8	469.9
79		250.6	507.0	268.9	468.2	296.2	445.1	304.2	436.2	317.9	422.4
80	Hemichordata	504.2	537.6	504.1	525.8	504.1	517.5	504.1	515.5	504.1	511.4
81		378.5	585.8	441.9	575.7	492.5	578.8	497.3	576.4	526.9	588.8
82	Protostomia	598.0	626.4	605.5	628.4	617.6	637.8	624.1	640.3	635.3	653.5
83		582.7	616.2	591.8	618.1	603.6	626.5	609.7	628.6	621.2	640.5
84		570.0	605.7	578.4	605.9	590.2	613.5	595.6	615.6	605.7	625.4
85	Annelids-Molluscs	552.3	586.1	559.2	585.3	567.2	588.8	570.6	590.3	577.4	595.1
86	Capitellid-Polychete-Leech	476.3	548.1	476.3	536.4	476.3	528.3	476.3	526.4	476.3	517.5
87		398.5	536.0	421.0	517.4	435.6	507.6	439.3	505.6	439.2	493.5
88		310.5	501.1	320.7	469.9	362.9	465.2	371.9	462.2	384.0	446.6
89		201.6	487.0	248.1	452.6	265.9	417.3	272.5	401.6	295.2	379.2
90	Mollusca	538.4	549.6	540.8	549.8	543.4	549.9	544.3	550.0	545.8	550.3
91	Bivalve-Gastropod	530.0	539.1	530.0	536.5	530.0	534.2	530.0	533.7	530.0	532.6
92	Gastropoda	470.0	508.3	470.1	497.9	470.0	487.4	470.0	484.6	470.0	478.8
93		265.0	516.5	304.5	486.8	313.2	452.3	314.3	444.6	324.8	431.8
94		310.6	541.4	317.1	512.1	344.0	502.9	349.3	489.6	394.2	481.2
95		72.7	452.4	90.5	350.9	110.4	254.0	122.5	248.1	140.6	225.2
96	Ecdysozoa	577.8	613.2	585.3	613.2	594.3	618.4	599.6	620.5	608.8	628.9
97	Nematoda-Arthropoda	561.4	599.8	568.6	598.4	577.7	602.4	581.6	604.0	589.8	610.4
98	Lobopodia	545.1	582.8	551.8	580.3	558.4	581.2	561.8	582.9	568.5	587.0
99	Euarthropoda	530.8	559.4	534.8	555.5	538.9	554.5	540.1	554.3	543.3	556.2
100	Mandibulata	523.4	532.3	526.2	532.6	528.4	533.1	528.9	533.4	530.3	536.1
101	Pancrustacea	514.0	522.8	514.0	520.4	514.0	518.5	514.0	518.0	514.0	517.5
102	Copepoda-Branchiopoda	499.0	510.1	498.9	505.8	498.9	502.5	498.9	501.8	498.9	500.5
103		414.4	496.1	420.1	485.7	423.7	476.8	426.1	474.9	435.2	468.1
104		324.8	441.5	328.0	424.0	330.4	392.4	332.8	387.2	334.8	374.4
105	Eumetabola	305.3	396.8	305.3	378.9	305.2	352.7	305.3	347.2	305.3	335.8
106	Pcynogonida-other chelicertates	497.5	526.1	497.4	520.5	497.4	514.7	497.4	512.5	497.4	509.1
107	Acari-Arenacea	415.9	479.9	415.9	466.3	415.8	453.2	415.7	448.4	415.8	436.4

Note.− Posterior times are the 95% HPD intervals, estimated with MCMCTREE under the LG+Γ_4+F model, using the calibration strategy 1 and different partitioning schemes. 1P: The 203 proteins analysed as a single partition. 2P, 4P, 5P, 10P: The proteins are grouped into 2, 4, 5, 10 partitions according to their evolutionary rates. Node numbers are as in Figure 6.6.

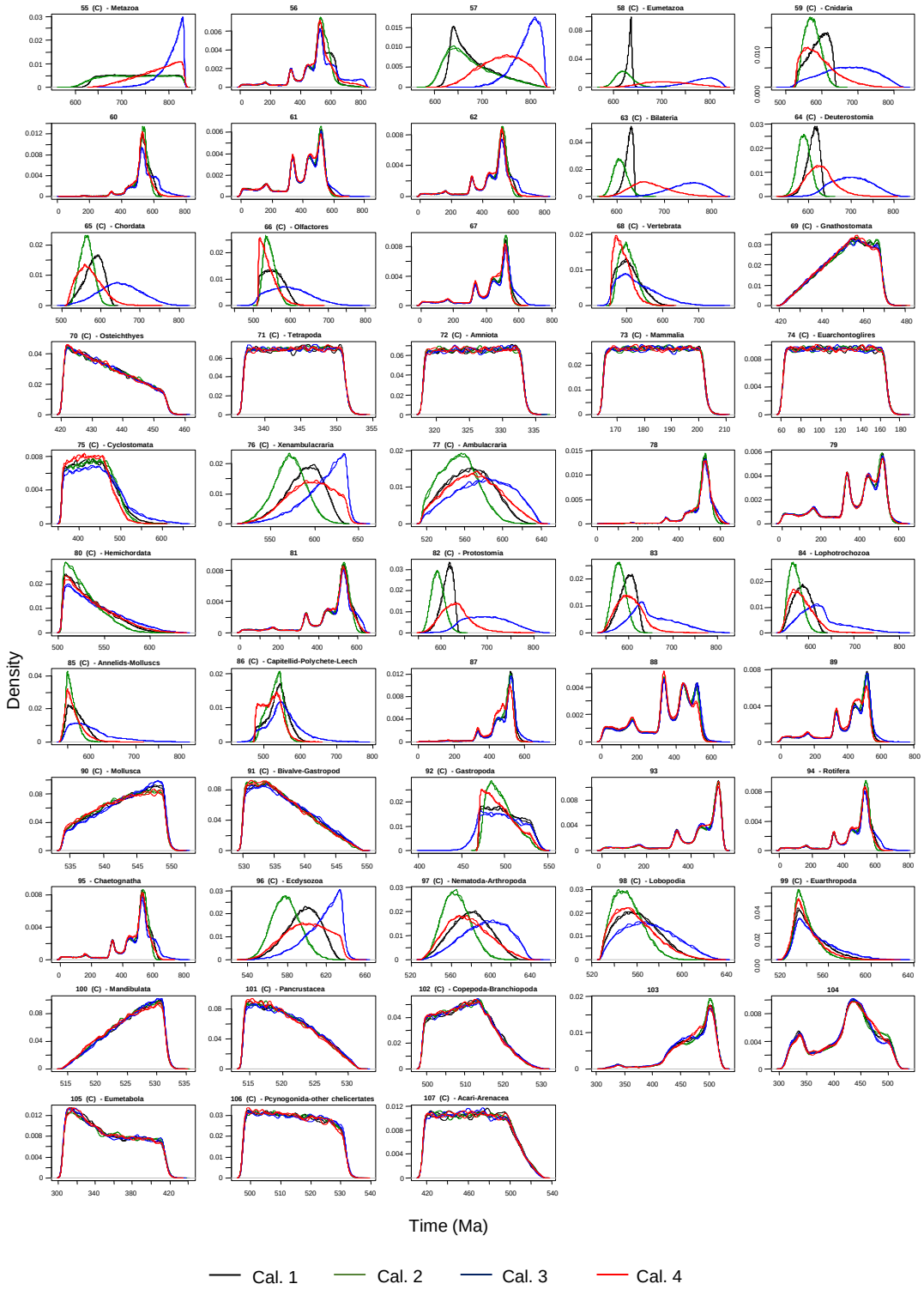

Figure D.1: Marginal prior densities of divergence times for all nodes in the tree under four different calibration strategies. For each strategy results from two MCMC runs are reported. Node numbers are as in Figure 6.6. Calibrated nodes are denoted with (C).

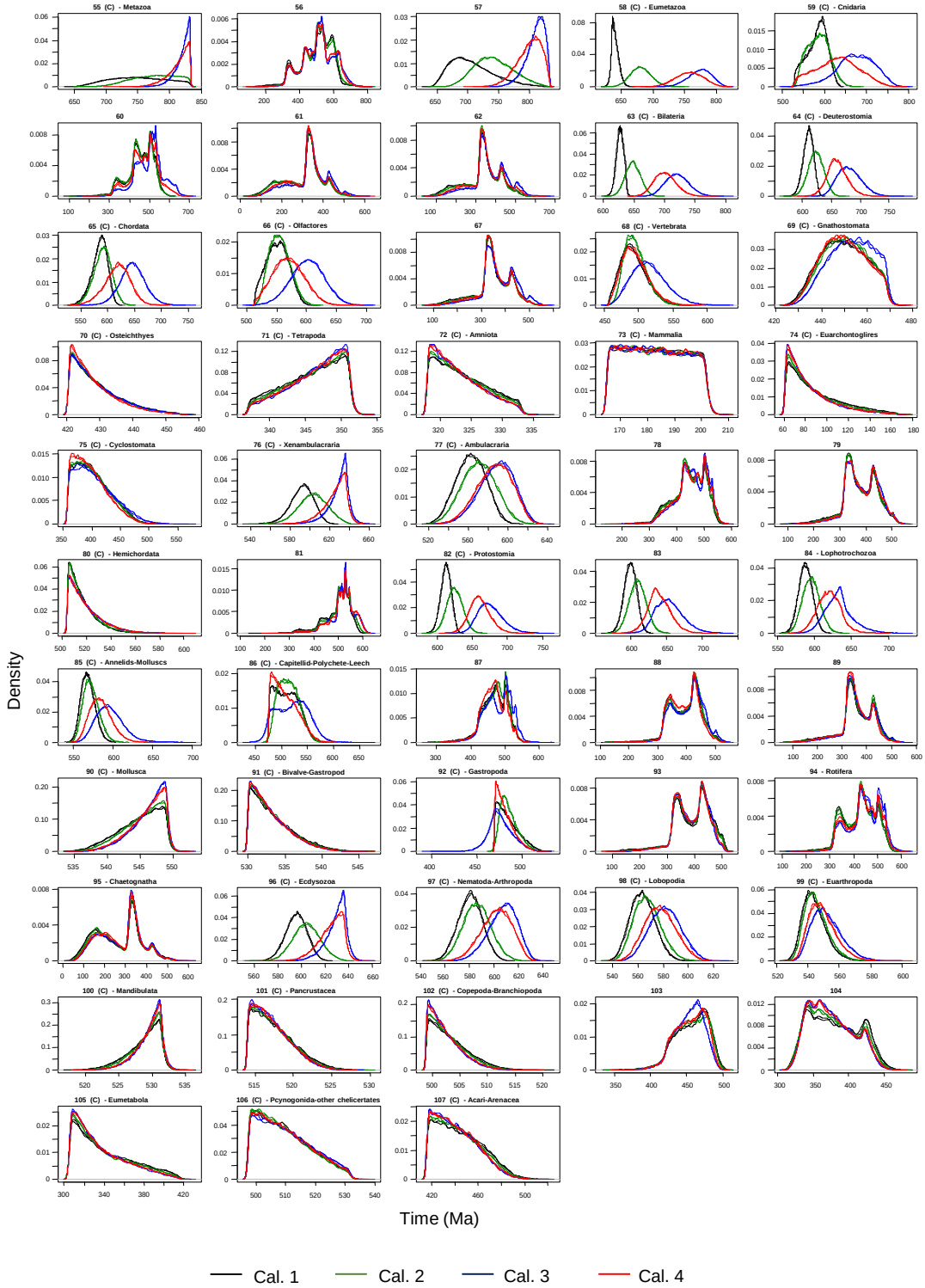

Figure D.2: Marginal posterior densities of divergence times for all nodes in the tree under four different calibration strategies. For each strategy results from two MCMC runs are reported. Node numbers are as in Figure 6.6. Calibrated nodes are denoted with (C).

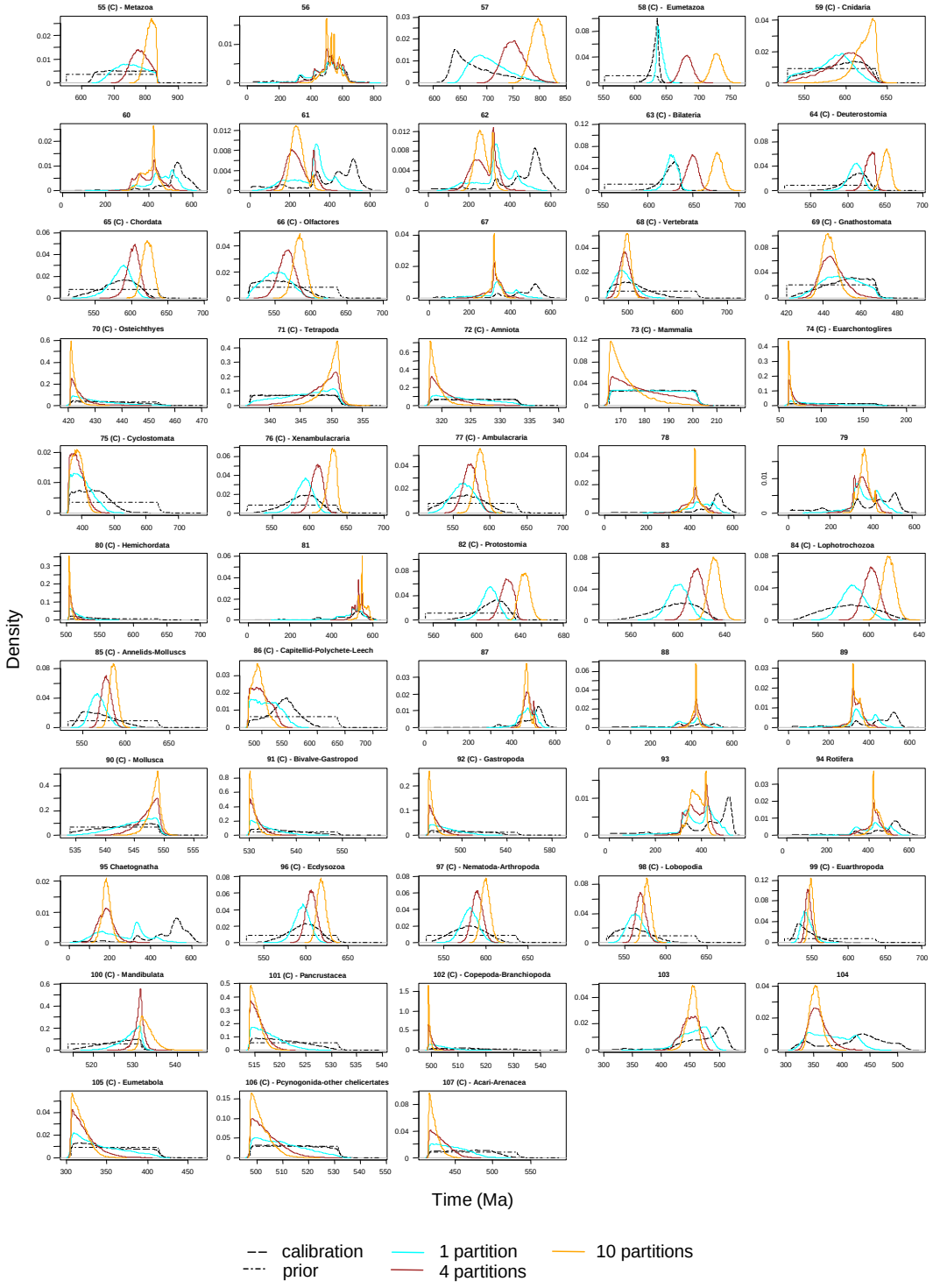

Figure D.3: Calibration, marginal prior and marginal posterior densities for various partitioning schemes under the calibration strategy 1. Node numbers are as in Figure 6.6. Calibrated nodes are denoted with (C).

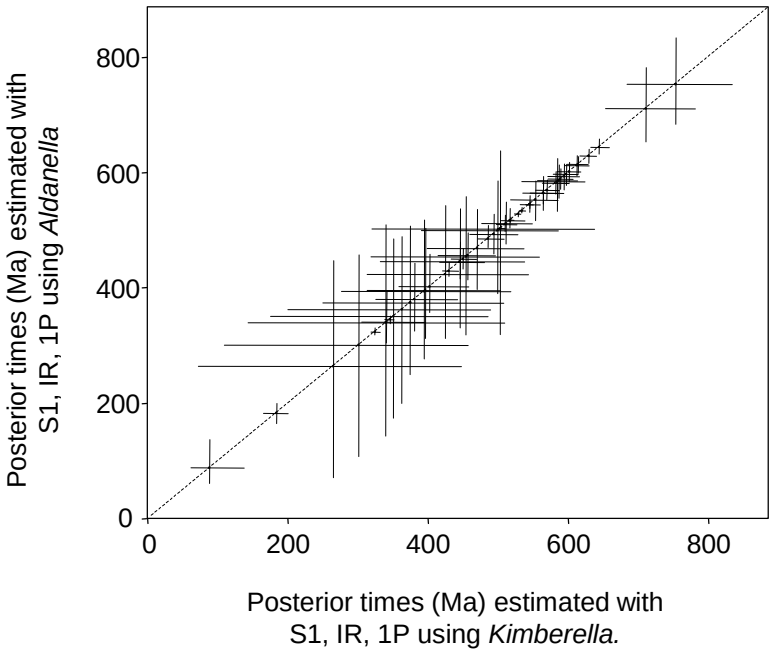

Figure D.4: Sensitivity of the time estimates to the fossil used to constrain basal clades in the metazoan phylogeny. The posterior mean times estimated using the fossil *Kimberella* to constrain the ages of the Metazoa, Eumetazoa, Bilateria and Protostomia clades are plotted against the estimates using the fossil *Aldanella*. The estimates are virtually identical.

References

Abramowitz M, Stegun IA. 1972. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover.

Allison A. 1954. Protection afforded by sickle-cell trait against subtertian malarial infection. Br Med J 1: 290–294.

Angelis K, dos Reis M. 2015. The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. Curr Zool.

Angelis K, dos Reis M, Yang Z. 2014. Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons. Mol Biol Evol 31(7):1902-1913.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950-958.

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 24:1219-1228.

Antcliffe JB, Callow RHT, Brasier MD. 2014. Giving the early fossil record of sponges a squeeze. Biol Rev 89:972-1004.

Arcila D, Pyron RA, Tyler JC, Orti G, Betancur-R R. 2015. An evaluation of fossil tip-dating versus node-age calibrations in tetraodontiform fishes (Teleostei: Percomorphaceae). Mol Phylogenet Evol 82:131-145.

Bajgain P, Richardson BA, Price JC, Cronn RC, Udall JA. 2011. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (Artemisia tridentata). BMC Genomics 12.

Barry D, Hartigan JA. 1987. Statistical analysis of hominoid molecular evolution. Stat Sci 2:191-210.

Barton NH, Keightley PD. 2002. Understanding quantitative genetic variation. Nat Rev Genet 3:11-21.

Bell CD, Soltis DE, Soltis PS. 2010. The age and diversification of the angiosperms re-revisited. Am J Bot 97:1296-1303.

Benton MJ, Donoghue PCJ. 2007. Paleontological evidence to date the tree of life. Mol Biol Evol 24:889-891.

Benton MJ, Donoghue PCJ, Asher RJ. 2009. Calibrating and constraining molecular clocks. In: Hedges SB, Kumar S, editors. The Timetree of Life. Oxford: Oxford University Press. p. 35-86.

Brandley MC, Schmitz A, Reeder TW. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst Biol 54:373-390.

Bromham L. 2011. The genome as a life-history character: why rate of molecular evolution varies between mammal species. Philos T Roy Soc B 366:2503-2513.

Bromham L, Penny D. 2003. The modern molecular clock. Nature Rev Genet 4:216-224.

Brown JM, Lemmon AR. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst Biol 56:643-655.

Budd GE. 2008. The earliest fossil record of the animals and its significance. Philos T Roy Soc B 363:1425-1434.

Budd GE, Jensen S. 2000. A critical reappraisal of the fossil record of the bilaterian phyla. Biol Rev 75:253-295.

Burbrink FT, Pyron RA. 2011. The impact of gene-tree/species-tree discordance on diversification-rate estimation. Evolution 65:1851-1861.

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol 25:1979-1994.

Burton A, Altman DG, Royston P, Holder RL. 2006. The design of simulation studies in medical statistics. Stat Med 25:4279-4292.

Buschiazzo E, Ritland C, Bohlmann J, Ritland K. 2012. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. BMC Evol Biol 12:8.

Cannarozzi GM, Schneider A. 2012. Codon Evolution: Mechanisms and Models. New York: Oxford University Press.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-552.

Charlesworth B, Lande R, Slatkin M. 1982. A neo-Darwinian commentary on macro-evolution. Evolution 36.

Clarke JT, Warnock RC, Donoghue PC. 2011. Establishing a time-scale for plant evolution. New Phytol 192:266-301.

Cooper A, Fortey R. 1998. Evolutionary explosions and the phylogenetic fuse. Trends Ecol Evol 13:324-324.

Darwin C. 1859. On the Origin of Species. London: John Murray.

Donoghue PCJ, Benton MJ. 2007. Rocks and clocks: calibrating the Tree of Life using fossils and molecules. Trends Ecol Evol 22:424-431.

Doolittle RF, Blomback B. 1964. Amino-acid sequence investigations of fibrinopeptides from various mammals - evolutionary implications. Nature 202:147-152.

dos Reis M. 2015. How to calculate the non-synonymous to synonymous rate ratio of protein-coding coding genes under the Fisher-Wright mutation-selection framework. Biol Lett 11:20141031.

dos Reis M, Donoghue PCJ, Yang Z. 2014. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biol Lett 10.

dos Reis M, Hay AJ, Goldstein RA. 2009. Using non-homogeneous models of nucleotide substitution to identify host shift events: application to the origin of the 1918 'Spanish' influenza pandemic virus. J Mol Evol 69:333-345.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. P Roy Soc Lond B Biol 279:3491-3500.

dos Reis M, Thawornwattana Y, Angelis K, Telford M, Donoghue PY, Yang Z. 2015. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr Biol.

dos Reis M, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol 28:2161-2172.

dos Reis M, Yang Z. 2013a. The unbearable uncertainty of Bayesian divergence time estimation. J Syst Evol 51:30-43.

dos Reis M, Yang Z. 2013d. Why do more divergent sequences produce smaller nonsynonymous/synonymous rate ratios in pairwise sequence comparisons? Genetics 195:195-204.

dos Reis M, Zhu TQ, Yang Z. 2014. The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. Syst Biol 63:555-565.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. Plos Biol 4:699-710.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-1973.

Duchene S, Lanfear R, Ho SYW. 2014. The impact of calibration and clock-model choice on molecular estimates of divergence times. Mol Phylogenet Evol 78:277-289.

Dunn CW, Giribet G, Edgecombe GD, Hejnol A. 2014. Animal phylogeny and its evolutionary implications. Annu Rev Ecol Evol Syst 45:371-395.

Edwards SV, Beerli P. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54:1839-1854.

Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ. 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. Science 334:1091-1097.

Fisher RA. 1930. The Genetical Theory of Natural Selection. Oxford: The Clarendon Press.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic lissamphibia. PLoS One 7:e48990.

Frandsen PB, Calcott B, Mayer C, Lanfear R. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol Biol 15:13.

Garcia-Gil MR, Mikkonen M, Savolainen O. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in Pinus sylvestris. Mol Ecol 12:1195-1206.

Ge GT, Cowen L, Feng XC, Widmer G. 2008. Protein coding gene nucleotide substitution pattern in the apicomplexan protozoa Cryptosporidium parvum and Cryptosporidium hominis. Comp Funct Genomics 2008:6 pages.

Gelman A, Rubin D. 1992. Inference from iterative simulation using multiple sequences. Stat Sci 7:457-511.

Gillespie JH, Langley CH. 1979. Are evolutionary rates really variable? J Mol Evol:27-34.

Gladieux P, Devier B, Aguileta G, Cruaud C, Giraud T. 2013. Purifying selection after episodes of recurrent adaptive diversification in fungal pathogens. Infect Genet Evol 17:123-131.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725-736.

Goldstein RA, Pollard ST, Shah SD, Pollock DD. 2015. Nonadaptive Amino Acid Convergence Rates Decrease over Time. Mol Biol Evol 32:1373-1381.

Goswami A, Upchurch P. 2010. The dating game: a reply to Heads (2010). Zool Scr 39:406-409.

Graves CJ, Ros VID, Stevenson B, Sniegowski PD, Brisson D. 2013. Natural selection promotes antigenic evolvability. Plos Pathog 9.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15:910-917.

Hasegawa M, Kishino H, Yano TA. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. J Mol Evol 22:160-174.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

Heads M. 2005. Dating nodes on molecular phylogenies: a critique of molecular biogeography. Cladistics 21:62-78.

Hedrick P. 2011. Genetics of Populations. Sudbury, MA: Jones & Bartlett Publishers.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol 27:570-580.

Ho SYW. 2014. The changing face of the molecular evolutionary clock. Trends Ecol Evol 29:496-503.

Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst Biol 58:367-380.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. Theoret Popul Biol:183-201.

Huelsenbeck J. 1995a. Performance of phylogenetic methods in simulation. Sys Bio 44:17-48.

Huelsenbeck J. 1995b. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol Biol Evol 12:843-849.

Huelsenbeck JP, Dyer KA. 2004. Bayesian estimation of positively selected sites. J Mol Evol 58:661-672.

Huelsenbeck JP, Jain S, Frost SWD, Pond SLK. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. P Natl Acad Sci USA 103:6263-6268.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. Nature 335:167-170.

Hughes AL, Ota T, Nei M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class-I major-histocompatibility-complex molecules. Mol Biol Evol 7:515-524.

Ina Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol 40:190-226.

Inoue J, Donoghue PCJ, Yang ZH. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. Syst Biol 59:74-89.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320-1331.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. Mammalian Protein Metabolism, edited by H. N. Munro Academic Press, New York:21-123.

Kainer D, Lanfear R. 2015. The effects of partitioning on phylogenetic inference. Mol Biol Evol 32:1611-1627.

Kendall DG. 1948. On the generalized birth-and-death process. Ann of Math Stat 19:1-15.

Kent WJ. 2002. BLAT - The BLAST-like alignment tool. Genome Res 12:656-664.

Kimura M. 1968. Evolutionary rate at molecular level. Nature 217:624-626.

Kimura M. 1969. The rate of molecular evolution considered from the standpoint of poulation genetics. Proc Natl Acad Sci USA 63:1181–1188.

Kimura M, Ohta T. 1971. Protein polymorphism as a phase of molecular evolution. Nature 229:467–469.

King JL, Jukes TH. 1969. Non-Darwinian evolution. Science 164:788-798.

Kingman JF. 1982a. The coalescent. Stoch Process Their Appl:235–248.

Kingman JF. 1982b. On the genealogy of large populations. J Appl Prob:27-43.

Kingsolver JG, Hoekstra HE, Hoekstra JM, Berrigan D, Vignieri SN, Hill CE, Hoang A, Gibert P, Beerli P. 2001. The strength of phenotypic selection in natural populations. Am Nat 157:245-261.

Kishino H, Thorne JL, Bruno WJ. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol 18:352-361.

Knowles LL, Kubatko LS. 2010. Estimating Species Trees: Practical and Theoretical Aspects. Hoboken, N.J.: Wiley-Blackwell.

Kondrashov AS, Turelli M. 1992. Deleterious mutations, apparent stabilizing selection and the maintenance of quantitative variation. Genetics 132:603-618.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. Protein Eng 8:641-645.

Koshi JM, Goldstein RA. 1998. Models of natural mutations including site heterogeneity. Proteins 32:289-295.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56:17-24.

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. Nature 392:917-920.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-132.

Lanfear R, Calcott B, Ho SY, Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695-1701.

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol 14:82.

Langergraber KE, Prufer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, et al. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc Natl Acad Sci USA 109:15716-15721.

Langley CH, Fitch WM. 1974. Examination of constancy of rate of molecular evolution. J Mol Evol 3:161-177.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286-2288.

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

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol 62:611-615.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B Biol Sci 363:3965-3976.

Leache AD, Harris RB, Rannala B, Yang Z. 2014. The influence of gene flow on species tree estimation: A simulation study. Syst Biol 63:17-30.

Leavitt JR, Hiatt KD, Whiting MF, Song H. 2013. Searching for the optimal data partitioning strategy in mitochondrial phylogenomics: a phylogeny of Acridoidea (Insecta: Orthoptera: Caelifera) as a case study. Mol Phylogenet Evol 67:494-508.

Lee MSY, Soubrier J, Edgecombe GD. 2013. Rates of phenotypic and genomic evolution during the Cambrian explosion. Curr Biol 23:1889-1895.

Li C, Lu G, Orti G. 2008. Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Syst Biol 57:519-539.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96-99.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150-174.

Lindley DV, Phillips LD. 1976. Inference for a Bernoulli process (a Bayesian view). Am Stat 30:112-119.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542-2543.

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

Liu L, Yu LL, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53:320-328.

Love GD, Grosjean E, Stalvies C, Fike DA, Grotzinger JP, Bradley AS, Kelly AE, Bhatia M, Meredith W, Snape CE, et al. 2009. Fossil steroids record the appearance of Demospongiae during the Cryogenian period. Nature 457:718-722.

Loytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. P Natl Acad Sci USA 102:10557-10562.

Magallon S, Hilu KW, Quandt D. 2013a. Land plant evolutionary timeline: Gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. Am J Botany 100:556-573.

Magallon S, Hilu KW, Quandt D. 2013b. Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. Am J Bot 100:556-573.

Magallon SA. 2004. Dating lineages: Molecular and paleontological approaches to the temporal framework of clades. Int J Plant Sci 165:S7-S21.

Majerus M. 2008. Industrial melanism in the peppered moth, biston betularia: An excellent teaching example of Darwinian evolution in action. Evo Edu Outreach 2:63-74.

Maloof AC, Porter SM, Moore JL, Dudas FO, Bowring SA, Higgins JA, Fike DA, Eddy MP. 2010. The earliest Cambrian record of animals and ocean geochemical change. Geol Soc Am Bull 122:1731-1774.

Maloof AC, Rose CV, Beach R, Samuels BM, Calmet CC, Erwin DH, Poirier GR, Yao N, Simons FJ. 2010. Possible animal-body fossils in pre-Marinoan limestones from South Australia. Nat Geosci 3:653-659.

Margoliash E. 1963. Primary structure and evolution of cytochrome C. P Natl Acad Sci USA 50:672-679.

Mau B, Newton MA. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. J Comput Graph Stat 6:122-131.

Mau B, Newton MA, Larget B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1-12.

Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TLL, Stadler T, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. Science 334:521-524.

Messier W, Stewart CB. 1997. Episodic adaptive evolution of primate lysozymes. Nature 385:151-154.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. J Chem Phys 21:1087-1092.

Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kuck P, Ebersberger I, Walzl M, Pass G, Breuers S, et al. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol Biol Evol 27:2451-2464.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763-767.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219-236.

Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol 16:23-36.

Monit C, Goldstein RA, Towers G, Hue S. 2015. Positive selection analysis of overlapping reading frames is invalid. AIDS Res Hum Retroviruses 31:947.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715-724.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426.

Nei M, Xu P, Glazko G. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. P Natl Acad Sci USA 98:2497-2502.

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol Evol 16:358-364.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. Plos Biol 3:976-985.

Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol 20:1231-1239.

Nielsen R, Yang ZH. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9.

Nuzzo R. 2014. Scientific method: statistical errors. Nature 506:150-152.

Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. Syst Biol 53:47-67.

O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo ZX, Meng J, et al. 2013. Response to comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". Science 341:613.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96-98.

Oliver JC. 2013. Microevolutionary processes generate phylogenomic discordance at ancient divergences. Evolution 67:1823-1830.

Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571-581.

Pamilo P, Bianchi NO. 1993. Evolution of the Zfx and Zfy genes - rates and interdependence between the genes. Mol Biol Evol 10:271-281.

Pauling L, Itano H, Singer S, Wells I. 1949. Sickle cell anemia, a molecular disease. Science 110:543–548.

Pellino M, Hojsgaard D, Schmutzer T, Scholz U, Horandl E, Vogel H, Sharbel TF. 2013. Asexual genome evolution in the apomictic Ranunculus auricomus complex: examining the effects of hybridization and mutation accumulation. Mol Ecol 22:5908-5921.

Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, Dodgson J. 1980. The evolution of genes - the chicken preproinsulin gene. Cell 20:555-566.

Peterson KJ, Butterfield NJ. 2005. Origin of the Eumetazoa: Testing ecological predictions of molecular clocks against the Proterozoic fossil record. P Natl Acad Sci USA 102:9547-9552.

Peterson KJ, Cotton JA, Gehling JG, Pisani D. 2008. The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philos T Roy Soc B 363:1435-1443.

Peterson KJ, Lyons JB, Nowak KS, Takacs CM, Wargo MJ, McPeek MA. 2004. Estimating metazoan divergence times with a molecular clock. P Natl Acad Sci USA 101:6536-6541.

Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470:255-258.

Pond SK, Delport W, Muse SV, Scheffler K. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. Plos One 5.

Pond SLK, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28:3033-3043.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Poux C, Madsen O, Glos J, de Jong WW, Vences M. 2008. Molecular phylogeny and divergence times of Malagasy tenrecs: influence of data partitioning and taxon sampling on dating analyses. BMC Evol Biol 8:102.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304-311.

Rannala B, Yang ZH. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645-1656.

Rannala B, Yang ZH. 2013. Improved reversible jump algorithms for Bayesian species delimitation. Genetics 194:245-253.

Rannala B, Yang ZH. 2007. Inferring speciation times under an episodic molecular clock. Syst Biol 56:453-466.

Raup DM. 1972. Taxonomic diversity during Phanerozoic. Science 177:1065-1071.

Rice W, Salt G. 1988. Speciation via disruptive selection on habitat preference: experimental evidence. Am Nat 131:911–917.

Robert C, Casella G. 2011. A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. Stat Sci 26:102-115.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray DL, Rasnitsyn AP. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst Biol 61:973-999.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539-542.

Rota-Stabelli O, Daley AC, Pisani D. 2013. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for Ecdysozoan evolution. Curr Biol 23:392-398.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol 14:23.

Runnegar B. 1982. A molecular-clock date for the origin of the animal phyla. Lethaia 15:199-205.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13:745-753.

Scheffler K, Seoighe C. 2005. A Bayesian model comparison approach to inferring positive selection. Mol Biol Evol 22:2531-2540.

Schenekar T, Winkler KA, Troyer JL, Weiss S. 2011. Isolation and characterization of the CYP2D6 gene in felidae with comparison to other mammals. J Mol Evol 72:222-231.

Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23:7-9.

Smith J. 1962. Disruptive selection, polymorphism and sympatric speciation. Nature 195:60-62.

Sokolowski J, Banks C. 2010. Modelling and Simulation Fundamentals. Hoboken, New Jersey: Wiley.

Sperling EA, Robinson JM, Pisani D, Peterson KJ. 2010. Where's the glass? Biomarkers, molecular clocks and microRNAs suggest a 200 million year missing Precambrian fossil record of Siliceous sponge spicules. Geobiology 8:24-36.

Springer MS, Amrine HM, Burk A, Stanhope MJ. 1999. Additional support for Afrotheria and Paenungulata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition. Syst Biol 48:65-75.

Steiper ME, Young NM. 2006. Primate molecular divergence dates. Mol Phylogenet Evol 41:384-394.

Stigler SM. 1986. The History of Statistics : The Measurement of Uncertainty Before 1900. Cambridge, Mass, London: Belknap Press of Harvard University Press.

Strugnell J, Norman M, Jackson J, Drummond AJ, Cooper A. 2005. Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) using a multigene approach; the effect of data partitioning on resolving phylogenies in a Bayesian framework. Mol Phylogenet Evol 37:426-441.

Sullivan J, Swofford DL. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst Biol 50:723-729.

Swanson WJ, Zhang ZH, Wolfner MF, Aquadro CF. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. P Natl Acad Sci USA 98:2509-2514.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437-460.

Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48:198-221.

Tamuri AU, dos Reis M, Hay AJ, Goldstein RA. 2009. Identifying changes in selective constraints: Host shifts in Influenza. Plos Comput Biol 5.

Thorne JL, Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst Biol 51:689-702.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15:1647-1657.

Voloch CM, Schrago CG. 2012. Impact of the partitioning scheme on divergence times inferred from Mammalian genomic data sets. Evol Bioinform Online 8:207-218.

Walters JR, Harrison RG. 2010. Combined EST and proteomic analysis identifies rapidly evolving seminal fluid proteins in Heliconius butterflies. Mol Biol Evol 27:2000-2013.

Wang DYC, Kumar S, Hedges SB. 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. P Roy Soc Lond B Bio 266:163-171.

Wang TC, Chen FC. 2013. The evolutionary landscape of the Mycobacterium tuberculosis genome. Gene 518:187-193.

Ward PS, Brady SG, Fisher BL, Schultz TR. 2010. Phylogeny and biogeography of dolichoderine ants: effects of data partitioning and relict taxa on historical inference. Syst Biol 59:342-362.

Warnock RCM, Parham JF, Joyce WG, Lyson TR, Donoghue PCJ. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. P Roy Soc Lond B Bio 282.

Warnock RCM, Yang ZH, Donoghue PCJ. 2012. Exploring uncertainty in the calibration of the molecular clock. Biol Lett 8:156-159.

Wong WSW, Yang ZH, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041-1051.

Wray GA, Levinton JS, Shapiro LH. 1996. Molecular evidence for deep precambrian divergences among metazoan phyla. Science 274:568-573.

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S, et al. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc Natl Acad Sci USA 109:17519-17524.

Yang Z. 1996a. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367-372.

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. Curr Zool.

Yang Z. 1994a. Estimating the pattern of nucleotide substitution. J Mol Evol 39:105-111.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573.

Yang Z. 1996b. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. J Mol Evol 42:587-596.

Yang Z. 1994c. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-314.

Yang Z. 2014. Molecular Evolution: A Statistical Approach. Oxford: Oxford University Press.

Yang Z. 2006. On the varied pattern of evolution of 2 fungal genomes: A critique of Hughes and Friedman. Mol Biol Evol 23:2279-2282.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496-503.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908-917.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32-43.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409-418.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000a. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000b. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600-1611.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23:212-226.

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

Yang Z, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. P Natl Acad Sci USA 107:9264-9269.

Yang Z, Rannala B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. Mol Biol Evol 31:3125-3135.

Yang Z, Roberts D. 1995. On the use of nucleic-acid sequences to infer early branchings in the tree of life. Mol Biol Evol 12:451-458.

Yang Z, Rodriguez CE. 2013. Searching for efficient Markov chain Monte Carlo proposal kernels. P Natl Acad Sci USA 110:19307-19312.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107-1118.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, et al. 2014. Three keys to the radiation of angiosperms into freezing environments. Nature 506:89-92.

Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat Commun 5:4956.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50:56-68.

Zhang JZ, Nielsen R, Yang ZH. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472-2479.

Zhu TQ, Dos Reis M, Yang ZH. 2015. Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. Syst Biol 64:267-280.

Zuckerkandl E, Pauling L. 1965. Evolving Genes and Proteins. In: Bryson V, Vogel HJ, editors. Evolutionary Divergence and Convergence in Proteins. New York: Academic Press. p. 97-166.

Zuckerkandl E, Pauling L. 1962. Horizons in Biochemistry. In: Kasha M, Pullman B, editors. Molecular Disease, Evolution, and Genetic Heterogeneity. New York: Academic Press. p. 189-225.