Bayesian molecular phylogenetics: estimation of divergence dates and hypothesis testing

Stéphane Aris-Brosou

University College London

Department of Biology

2002

A thesis submitted to the University of London

for the degree of Doctor of Philosophy.

Abstract

Bayesian molecular phylogenetics:

estimation of divergence dates and hypothesis testing

Stéphane Aris-Brosou
University College London

D. Phil. Thesis
University of London

With the advent of automated sequencing, sequence data are now available to help us understand the functioning of our genome, as well as its history. To date, powerful methods such as maximum likelihood have been used to estimate its mode and tempo of evolution and its branching pattern. However, these methods appear to have some limitations. The purpose of this thesis is to examine these issues in light of Bayesian modelling, taking advantage of some recent advances in Bayesian computation.

Firstly, Bayesian methods to estimate divergence dates when rates of evolution vary from lineage to lineages are extended and compared. The power of the technique is demonstrated by analysing twenty-two genes sampled across the metazoans to test the Cambrian explosion hypothesis. While the molecular clock gives divergence dates at least twice as old as those indicated by the fossil records, it is shown (i) that modelling rate change gives results consistent with the fossils, (ii) that this improves dramatically the fit to the data and (iii) that these results are not dependent on the choice of a specific model of rate change. Results from this analysis support a molecular explosion of the metazoans about 600 million years (MY) ago, i.e. only some 50 MY before the morphological Cambrian explosion.

Secondly, two new Bayesian tests of phylogenetic trees are developed. The first aims at selecting the correct tree, while the second constructs confidence sets of trees. Two other tests are also developed, in the frequentist framework. Based on *p*-values adjusted for multiple comparisons, they are built to match their Bayesian counterparts. These four new tests are compared with previous tests. Their sensitivity to model misspecification and the problem of regions is discussed.

Finally, some extensions to the models examined are made to estimate divergence dates from data of multiple genes, and to detect positive selection.

Contents

List of tables

List of figures

Probability theory is nothing but common sense reduced to calculation.

(Laplace)

Historical truth is not what has happened; it is what we judge to have happened.

(J.L.Borges)

Grey, my dear child, is any theory
And green is the golden tree of life.
(Goethe – Faust I)

Acknowledgments

I would like to thank my supervisor, Professor Ziheng Yang, for the absolute liberty he has given me. His clear views on many issues, his patience and his purchase of fast computers helped a lot. The financial support of the Biotechnological and Biological Sciences Research Council is gratefully acknowledged.

I would also like to thank Professor Hirohisa Kishino for hosting me at the University of Tokyo (東大) during a short visit funded by the Ministry of Education, Science, Sports and Culture of Japan (Monbusho、文部省). This visit gave me the opportunity to broaden my views on the two main subjects of this thesis.

David Balding and Nick Goldman helped to improve the manuscript and correct some of the errors, in particular with respect to Chapter IV. Many thanks also to Joe Bielawski and Lounès Chikhi for discussions. This list would not be complete without a mention of my parents, who tried to understand and sustain the enthusiasm that kept me going throughout.

Ceci est pour vous.

Introduction

At the dawn of the XVIIIth century, it was still believed that living beings had been created by a Divine act, estimated to have occurred some 6,000 years ago according to the Bible. Driven by diverse considerations, animals were ranked and pigeonholed hierarchically, hereby generating a *system*. The observed diversity in terms of categories was explained by the concept of transformism where a category could metamorphose into another one. A scientific revolution marked the following century where the success of the Newtonian mechanics incited Kant (1724-1804) and Laplace (1749-1827) to introduce the notion of History in the Western cosmogony, but the concept did not infuse natural sciences immediately. The first slit to the creationist and transformist paradigms appeared with Lamarck (1744-1829), who postulated that living forms were changing over time, evolved, following a linear succession in which organisms were ordered from the most simple up to the most complex, along what he called a *scala naturae*. The key point is that he also postulated a mechanism, the theory of acquired characters, by which species were evolving. Lamarck's ideas were not particularly welcomed, mainly because of the influence at that time in France of Cuvier (1769-1832), said to have killed him a second time when he gave Lamarck's funeral oration (e.g. Gould, 2001, p.117). The linear arrangement of species was also criticised by von Baer (1792-1876) who, among others, was classifying animals according to a number of developmental types deriving from a common one. It is mainly with Darwin (1809-1882) and the publication of the *Origin of species* (Darwin, 1859), that it became possible to link the different species historically, that is to relate them according to a tree of organisms, if not yet a tree of life (Figure 0.1). Although Darwin seems to have hesitated to derive all the represented taxa from a unique ancestor, the idea was present: Pre-Darwinian biologists interpreted a natural system; with the *Origin of Species*, natural systematic categories exist because

organisms descend from a common ancestor. From this time on, such trees or phylogenies have been largely represented, either as abstract graphs (Figure 0.1) or under the more poetic shape of an actual tree as in Haeckel (1866) (Figure 0.2).

Phylogenetics can be defined as the study of patterns of evolution in relation to the history of organisms, in their ancestor to descendent relationship. The objectives are clearly defined, as this extract shows: "*As soon as phylogenetic considerations were added to systematics, three new questions arose. What are the phylogenetic relationships, or which stem branched off where? When in geological or relative evolutionary time did a given branching take place? How rapid was the evolutionary rate of a given line in a given time period?*" (Sokal and Sneath, 1963, p. 24). Two possible approaches can be undertaken. In the first place, this reconstruction can be made directly from the fossil records. This is viewed as an ideal (Nei and Kumar, 2000, p. 3), and is difficult to achieve indeed. Since Darwin (*ibid.*, Chapter IX: "On the imperfection of the geological record"), the paleontological evidence is held to be fragmentary and incomplete (Fortey et al., 1996, but see Baumiller, 1999) so that history must be reconstructed from indirect approaches. Until recently, the alternative was to resort to comparative biology, and in particular, since Cuvier, to comparative anatomy and physiology. The rise of genetics permitted further developments.

Deoxyribonucleic acid (DNA), and for some viruses ribonucleic acid (RNA), form the support of hereditary information. The rules governing its transmission, known as Mendel's laws, were established in the late XIXth – early XXth centuries, but it is only in the last three decades that advances in molecular biology, both in terms of concepts (Kimura's neutral theory was published in 1969) and techniques (the entire human genome was obtained in 2001 by shotgun sequencing and automated gene-sequencing machines), allowed us to understand the mechanisms governing the evolution of these

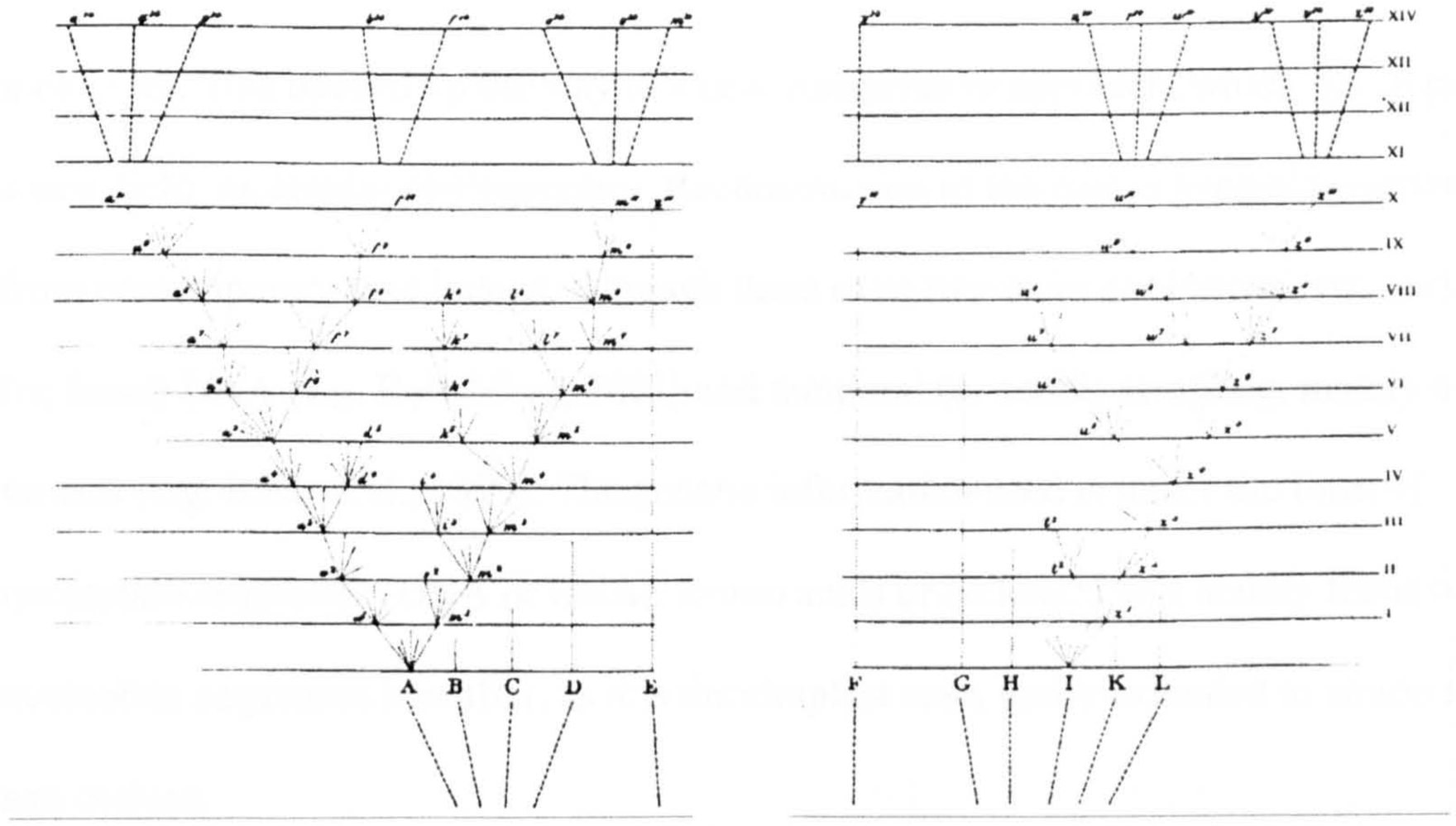


Figure 0.1. The only figure present in Darwin's book *The Origin of species*, and one of the first phylogenetic trees published.

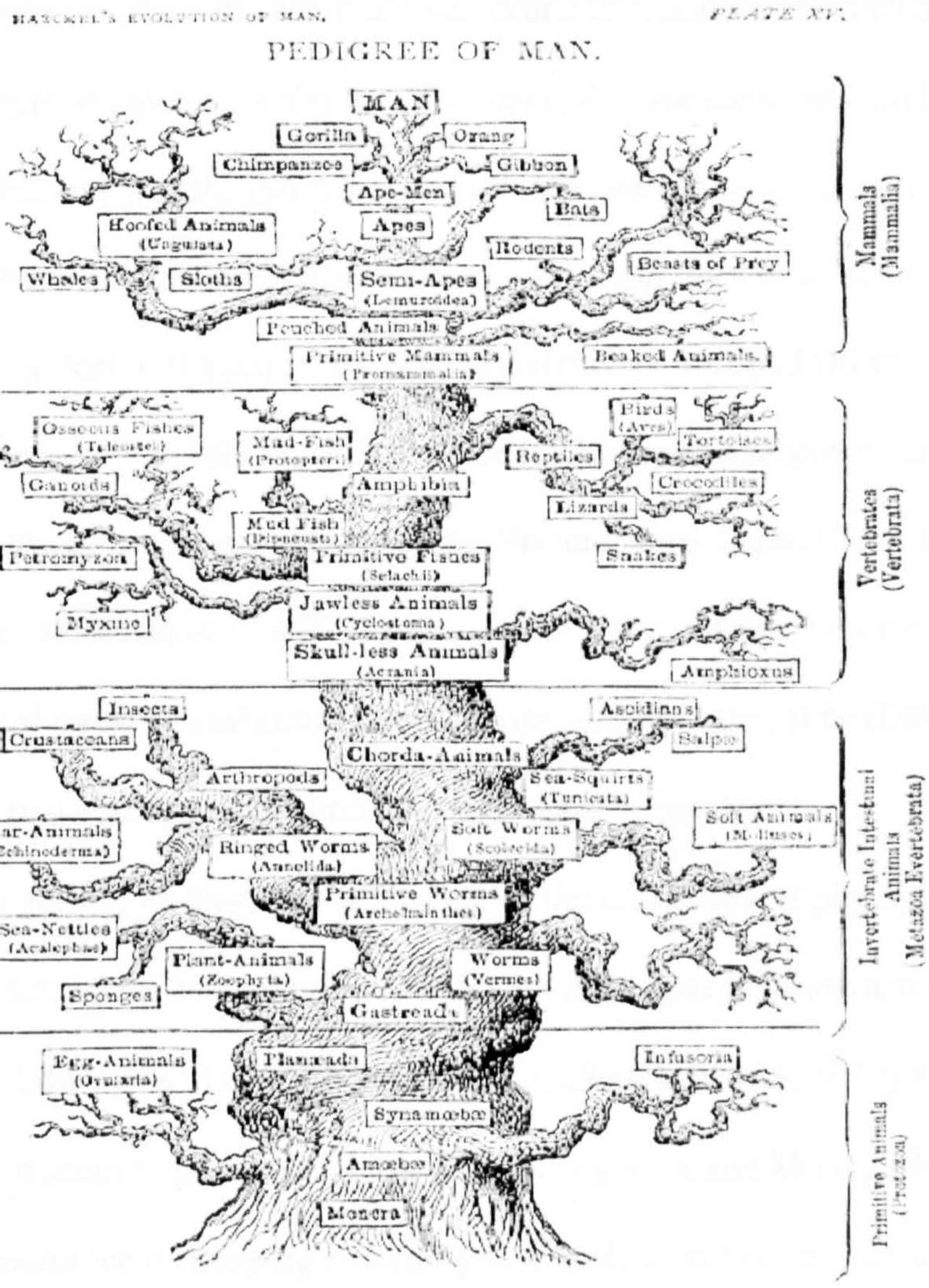


Figure 0.2. The last plate to be found in Haeckel's book.

molecules. This opened up the way to a new comparative approach, which developed into a new field: molecular phylogenetics. Reconstruction of the past is generally carried out from contemporary taxa indeed, although there exist two main exceptions, viz. ancient (or fossil) DNA (e.g. Relethford, 2001) and temporal (or serial) sampling, mainly of viruses (e.g. Bush et al., 1999). The genetic information used is under the form of nucleotide sequences (DNA or RNA), amino acids or codons. I will mainly focus on nucleotide sequences hereafter, as it is the simplest case, easily extended to amino acids and codons.

Until recently, reconstruction methods were essentially divided between two schools. In the "phylogenetic school" (Hennig, 1966), the branching pattern is obtained on the basis of the identification of shared derived characters, and reconstruction mainly relies on the principle of parsimony. On the other hand, the tree reconstructed by the "phenetic school" (Sokal and Sneath, 1963) is based on measurements of similarity. These approaches have been efficient in determining the main lines of evolutionary history, but they difficultly deal with homoplasy, and confidence in the obtained reconstructions is difficult to assess. The problem is most evident when such analyses return several trees that explain the data with equal plausibility. Because each character evolves in a highly dimensional "morphospace", statistical models are difficult to set up for the evolution of morphological traits, so that conflicting theories (models) cannot be disentangled easily. The genetic material appeared almost at once easier to process.

The first genetic studies mainly focused on the estimation of phylogenetic trees from continuous data: gene frequencies within populations (Cavalli-Sforza and Edwards, 1964, 1966, 1967; Edwards and Cavalli-Sforza, 1963, 1964; Edwards, 1970). Gene-splitting was modelled according to a pure birth process (e.g. Cox and Miller, 1965, p.156), while gene frequencies were changing according to a random walk (e.g. Cox and Miller, 1965,

p.46). Although these two models are quite simple, they caused overwhelming computational problems (Edwards, 1970; Thompson, 1975). This work remained relatively obscure until Felsenstein (1973, 1981) considered a discrete type of data, nucleotide sequence data, to which he applied the method of maximum likelihood developed by R.A. Fisher (1890-1962). The tree and the branch lengths are now parameters that can be directly estimated from the data. Despite the progress made since these beginnings, two issues are still debated. Firstly, one of the objectives of phylogenetics is to estimate divergence times (Adachi and Hasegawa, 1996, p. 32). However, since rates and times are not identifiable, this is not directly possible. Secondly, assessing the confidence of estimated trees is still an active area of research (e.g. Holmes, 1999; Shimodaira and Hasegawa, 1999; Goldman et al., 2000; Nei and Kumar, 2000, pp. 165-186; Ewens and Grant, 2001, pp. 416-421). The reason is that the tree is not a parameter in the common acceptance of the term (Yang et al., 1995), and a particular bifurcating tree cannot be reduced to a special case of another one. Added to the problems of multiple comparisons and selection bias, testing phylogenetic trees is not trivial. However, these limitations do by no means toll the bell of statistical methods to favour non-probabilistic ones evoked earlier (for a review of these methods, see Swofford et al., 1996).

This thesis addresses these two issues of estimating divergence dates and assessing confidence in estimated tree topologies mainly in a framework based on the theorem of Reverend T. Bayes (1702-1761; buried in the Bunhill Fields cemetery, London). The structure of the thesis is as follows:

Chapter I provides an overview of maximum likelihood methods to estimate divergence dates and test phylogenetic hypotheses.

Chapter II is devoted to the presentation of Bayesian models and techniques to estimate divergence dates. The models are compared with their maximum likelihood counterparts by the study of small data sets, and their sensitivity is evaluated in a simulation study.

Chapter III gives an application of the methods presented in Chapter II, re-evaluating the molecular evidence for a Precambrian explosive diversification of multicellular animals.

Chapter IV introduces the concepts of model selection and of construction of confidence sets of trees. Four new tests are developed and compared: two are frequentists and two are Bayesians.

Chapter V presents some extensions of the general approach to take into account model uncertainty, models incorporating multiple genes and methods to detect positively selected lineages and selection of models of positive selection.

A computer program is presented in Appendix 1, illustrated with a short analysis of the timescale of the evolution of the rodents. This program can be used to address the aforementioned issues. Appendix 2 contains the accession numbers of the sequences used in Chapter III. Lastly, Appendix 3 amends the implementation of the two new frequentist tests presented in Chapter IV.

Chapter I

Likelihood methods in phylogenetics

Likelihood is a central concept in statistical inference. When an explicit model is formulated, this quantity is intuitively appealing as it is proportional to the probability of the data given a parameter, which may be a vector. In the simplest case where the parameter is real-valued, its estimate is the value that maximises the likelihood, assuming the model is correct. The use of likelihood in this context is referred to as the method of maximum likelihood (ML). While difficulties may arise when the model includes unwanted or "nuisance" parameters (Edwards, 1992, p.109, proposes to restructure the model), choosing the ML method against other methods of statistical inference, such as the least squares method, has some theoretical justifications that I expose briefly.

Estimators are generally easy to construct, and can be distinguished according to a number of properties. Ideally, we want an estimator that takes the true value θ_0 of the parameter θ of interest with probability one, irrespective of the sample realisation. This ideal estimator can be described with respect to its first two moments: (i) $E(\theta) = \theta_0$ and (ii) $\text{var}(\theta) = 0$. For a finite sample size n, property (ii) cannot be emulated, but as n goes to infinity some estimators can achieve it indeed. This leads to the distinction between finite and asymptotic sample properties.

Five properties characterise finite sample sizes. First, an estimator is said to be *unbiased* if its sampling distribution has an expectation equal to the true parameter (see property (i) above). Second, for two unbiased estimators $\mathscr{E}_1$ and $\mathscr{E}_2$ of θ, $\mathscr{E}_1$ is said to be more *efficient* than $\mathscr{E}_2$ if $\text{var}(\mathscr{E}_1) \leq \text{var}(\mathscr{E}_2)$. However, an estimator better than a terrible estimator is not necessarily a good estimator. Conversely, the existence of a lower bound is of interest: an estimator with a variance that no other estimator can better is said to be fully efficient. Assuming that the Fisher information for a sample of size n (expectation of minus the second derivative of the log of the density of the sample with respect to the parameter θ), $\mathscr{I}_n(\theta)$, exists and is strictly positive for all θ, the variance of any unbiased

estimator cannot be smaller than inverse of $\mathcal{I}_n(\theta)$. The Cramér-Rao lower bound can be derived from $\mathcal{I}_n^{-1}(\theta)$ for any estimator (Shao, 1999; p. 135). Third, fully efficient and unbiased estimators do not always exist, so that we may have to consider biased estimators and eventually compare them with unbiased or other biased estimators. The most widely used measure is the mean square error (MSE), defined as

$$\mathrm{E}\{(\mathcal{T}(\theta)-\theta)^2\} = \mathrm{var}[\mathcal{T}(\theta)] + \{\mathrm{E}[\mathcal{T}(\theta)]-\theta\}^2 \qquad \text{(I.1)}$$

that is, the sum of the variance and of the bias of the estimator. An estimator $\mathcal{T}_1(\theta)$ is said to be a *minimum MSE estimator* of θ if $\mathrm{MSE}\{\mathcal{T}_1(\theta)\} \leq \mathrm{MSE}\{\mathcal{T}_2(\theta)\}$ for any other estimator $\mathcal{T}_2(\theta)$. Two properties of practical importance are *parameterisation invariance*, where the estimator is not altered by a one-to-one transformation, and *sufficiency*, which characterises statistics $\mathcal{T}$ (estimators are statistics) reducing the data with no loss of information (the conditional probability of the data X given $\mathcal{T}(\theta(X))$ is known, i.e. is not dependent on any population or parameter).

Three asymptotic properties play an important role in estimation theory. An estimator is *consistent* if it converges in probability to the true value when the sample size tends to infinity. If there exists a function f_n of the sample size n such that $f_n(\mathcal{T}_n(\theta)-\theta)$ is asymptotically distributed as $\mathrm{N}(0,\mathrm{var}_\infty[\theta])$ with $\mathrm{var}_\infty[\theta] \neq 0$, then $\mathcal{T}(\theta)$ is said to be *asymptotically normal*. A consistent and asymptotically normal estimator is said to be *asymptotically efficient* if $f_n(\mathcal{T}(\theta)-\theta)$ is asymptotically distributed as $\mathrm{N}(0,[\mathcal{I}_\infty(\theta)]^{-1})$, assuming $\mathcal{I}_\infty(\theta) \neq 0$. That is, the asymptotic variance equals the asymptotic Cramér-Rao lower bound.

Maximum likelihood estimators are known to be parameterisation-invariant, unbiased, fully efficient and sufficient. They are asymptotically consistent,

asymptotically normal and asymptotically efficient. These are very strong arguments in favour of the use of maximum likelihood, in particular in phylogenetics.

An important aspect of ML inference is that likelihood is based on an explicit model. ML methods are sometimes assumed to be robust to violations of the assumptions of the model (Swofford et al., 1996, p. 430), but when several models can be formulated, it is desirable to base inference on the model that best explains the data (but see Yang, 1997a). The likelihood framework can be used for model selection, and also to test hypotheses.

Statistical inference is usually divided into two areas: point estimation, and hypothesis testing (with interval estimation often considered to bridge the gap). This subdivision is also clear in likelihood-based phylogenetics. I review in this chapter how it has been applied to the modelling of evolutionary processes, which aims at estimating parameters. Some of the assumptions of the model appear critical when the parameter of interest is the vector of divergence times. Therefore, different tests are reviewed.

I.1 – The likelihood of a tree and its computation

The idea of likelihood in phylogenetics was introduced by Cavalli-Sforza and Edwards (1967), but the concept only received a practical presentation a decade later with Felsenstein (1981). He proposed an algorithm, called the "pruning algorithm", to compute the likelihood of a tree for any number of taxa. The original presentation, for nucleotide data (Felsenstein, 1981), was later extended to amino acid sequences (Kishino et al., 1990; Adachi and Hasegawa, 1992) and to codon data (Goldman and Yang, 1994; Muse and Gaut, 1994). Here, I describe the general procedure applied to models of nucleotide substitution. This thesis focuses on this type of data, but these models are readily extended to accommodate the other data types. These models are based on

Markov processes discrete in space and continuous in time (e.g. Cox and Miller, 1965), and assume that all the sites are independently and identically distributed (iid). As rates of evolution are known to vary, both in time (among lineages) and in space (among sites), I present in the last subsection models of rate change in the likelihood framework.

I.1.a – The probability of a tree and its optimisation

Computation of the likelihood: the pruning algorithm

The probability of observing data x_h at a particular site h given a topology T and the parameters θ of a model of evolution is noted $p(x_h \mid T, \theta)$. This quantity is usually referred to as the probability of θ at this site, f_h, and is proportional to the sitewise likelihood at site h. It is efficiently calculated using the pruning algorithm (Felsenstein, 1981; see also Yang, 2000b). The computation of the probability of observing x_h under a given tree proceeds from a hypothetical root, which is not identifiable when time-reversible models are used (see I.1.b). The root is therefore located at a place convenient for computation, e.g. at an existing internal node, such as node i in Figure I.1.

In the case depicted in Figure I.1, node i has two direct descendents: nodes k and l. The conditional probability of observing data at the tips of the tree that are descendents of node i, $x_h(i)$, given that node i is in state x_i ($x_i \in$ {T, C, A, G}), noted $p_i(x_h(i) \mid x_i)$, is obtained by the product of two terms:

$$\sum_{x_k} p_{x_i x_k}(b_k) \cdot p_k(x_h(k) \mid x_k) \times \sum_{x_l} p_{x_i x_l}(b_l) \cdot p_l(x_h(l) \mid x_l) \qquad \text{(I.2)}$$

where b_k is the branch length or the time elapsed between nodes i and k. The first sum for instance represents the probability of node i in state x_i changing to state x_k during the interval b_k times the conditional probability of observing the data at nodes descending from node k given that node k is in state x_k. These conditional probabilities $p_i(x_h(i) \mid x_i)$ are computed recursively down the terminal taxa, also called leaves by mathematicians.

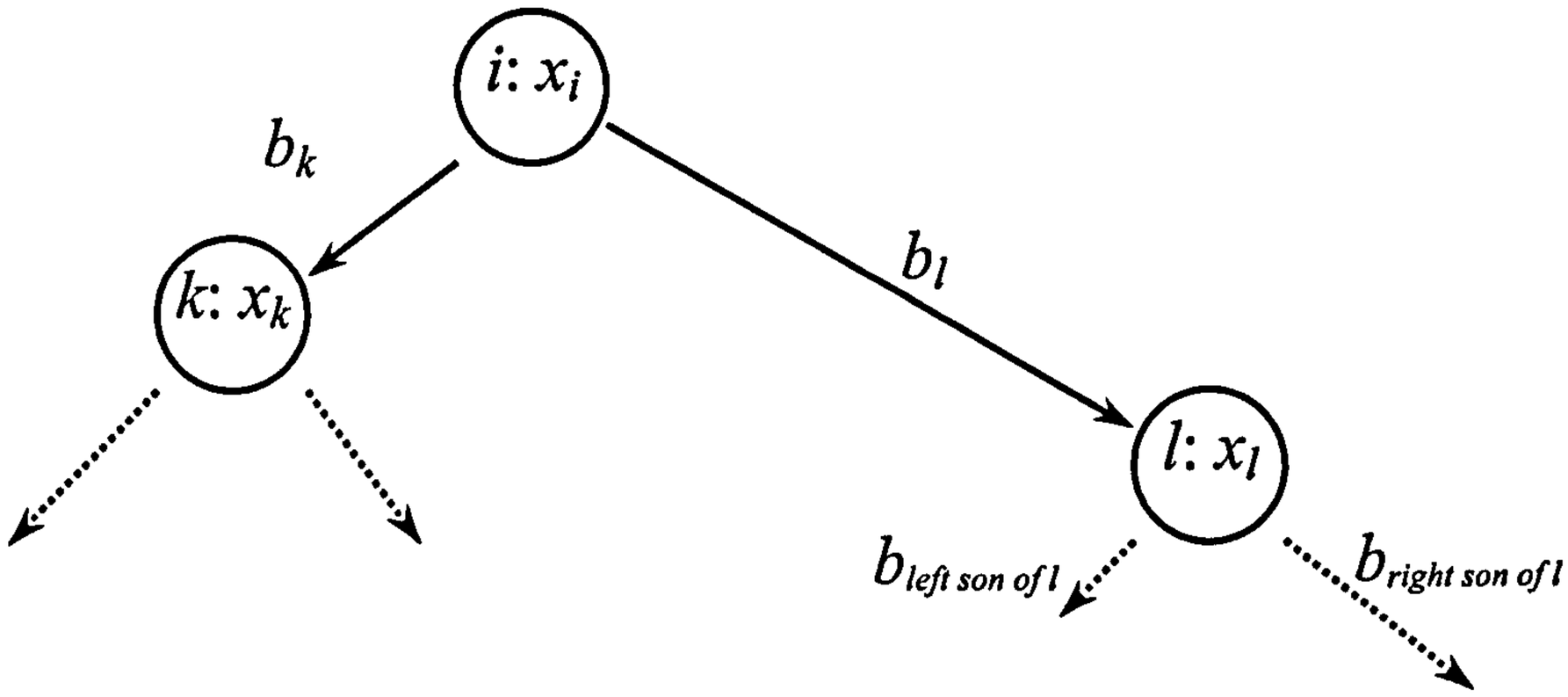


Figure I.1. Computation of the sitewise likelihood $\ell_i(\theta)$ at a node i in state x_i that has two descendents k and l, in state x_k and x_l, respectively (see text for details). The process is repeated recursively down to the leaves (broken arrows) to compute the site-probability of the tree from the internal node i.

When a leaf is reached, the probability of state x_k is one if x_k is the observed nucleotide; otherwise it is zero. If the data contains unidentified nucleotides (a given class of nucleotides such as a purine, or missing data), $p_i(x_h(i) \mid x_i)$ is set to one for each x_i compatible with the ambiguous data.

Starting the computation from the unobserved hypothetical root arbitrarily placed at the internal node i (Figure I.1), the unconditional probability of observing the data, f_h, is:

$$\sum_{x_i} \pi_{x_i} \cdot p_i(x_h(i) \mid x_i) \qquad \text{(I.3)}$$

where π_{x_i} is the prior probability of state x_i at the root i. Note that when a branch length is modified, such as the branch leading to the left son of l, $b_{\text{left son of } l}$ (Figure I.1), the only computations that need to be redone are those for the branches $b_{\text{left son of } l}$ up to the branch emanating from the "root", b_l. The number of computations can be large (Nei and Kumar, 2000, p.149), but the burden is generally not prohibitive and this algorithm can be applied to compute the likelihood of a tree of any size.

Finally, the likelihood of the whole sequence is obtained by multiplying the site likelihood values f_h's for each site position h as calculated above. This assumes that the sites are independently distributed. As these probabilities are usually very small, it is more convenient to work on a log-scale, in which case the sitewise log-likelihood values are added. This general algorithm is improved by compressing the original data set to a matrix of site patterns, on which the likelihood computations are done. The calculated value at each site pattern is then weighted by the observed frequency of the pattern.

The iid assumption may not be realistic. For instance, purines tend to follow purines and pyrimidines tend to follow pyrimidines (Barry and Hartigan, 1987b), and long-range autocorrelations have been reported (Peng et al., 1992). When sites are not iid, the model tends to become more complicated. Barry and Hartigan (1987b) gave an algorithm by which they demonstrated that the conditional probability of a T following a C is larger

than that of a G following a C for a given proportion of Cs in human mitochondrial DNA. Schoniger and von Haeseler (1994) proposed a model that substitutes independent and non-overlapping doublets of nucleotides while preserving autocorrelation of the doublets, and showed that neglecting autocorrelation tends to underestimate branch lengths. Yang (1995) and Felsenstein and Churchill (1996) described models with autocorrelated evolutionary rates in the likelihood framework.

Optimisation algorithms for real-valued parameters

Optimisation of the branch lengths b can be done by one of the following methods, either by a updating all the model parameters at the same time, or by updating parameters one by one. In the first case, imagine to simplify that there are only two parameters to optimise, such as two branch lengths: b_1 and b_2. Let the likelihood function be represented by a contour plot in a two-dimension plane. The algorithm starts from a random place in the plane, $B^{(0)} = \{b_1^{(0)}, b_2^{(0)}\}$ and a random direction, hereby defining a random initial line. Under some regularity conditions, the likelihood surface has a maximum along this line, found by a linear search algorithm. Hence, both branch lengths b_1 and b_2 are updated simultaneously, and take the value at the point $B^{(1)}$ that maximises the likelihood. The next step consists in searching the direction of maximum slope $\nabla^{(1)}$ on the likelihood surface at $B^{(1)}$ and the line $(B^{(1)}, \nabla^{(1)})$ is drawn. A similar linear search (maximisation) is carried out to obtain $B^{(2)}$. The process is repeated until some convergence criterion is reached. Gill et al. (1981) give more details about this algorithm, implemented in PAML (Yang, 1997b). Alternatively, one parameter at a time can be estimated by maximising the likelihood function for each parameter. The solution is obtained numerically by means of the Newton-Raphson method (Edwards, 1992, pp. 87-84;Adachi and Hasegawa, 1996, p. 40). This method takes advantage of the convexity of the likelihood function. If $\hat{\theta}$ is an estimator of θ, its first derivative is zero at $\hat{\theta}$: let $T(\theta)$

be the first derivative of the likelihood function with respect to θ; from an initial guess (or the t^{th} iteration) θ_t, $T(\hat{\theta})$ can be approximated around θ_t by $T(\theta_t) + (\hat{\theta} - \theta_t)\, dT(\theta_t)/d\theta$, so that the parameter value at the next iteration is chosen as $\theta_{t+1} = \theta_t - T(\theta_t)/[\, dT(\theta_t)/d\theta]$. When the algorithm is not convergent, a modified Newton algorithm proposes the next step as:

$$\theta_{t+1} = \theta_t - \alpha\, T(\theta_t) / [\, dT(\theta_t) / d\theta\,] \qquad \text{(I.4)}$$

where the step length $\alpha > 0$ is reduced repetitively until the likelihood at θ_{t+1} is not worse than the one at θ_t (Gill et al., 1981; Yang, 2000b).

Number of possible topologies and tree search strategies

Branch lengths are not the only parameters to optimise. The first aim of phylogenetics is to estimate the relationship linking different species (see Introduction). This relationship is expressed by means of a tree, which is singular for at least two reasons. First, its status is still debated: is it a random variable? (Cavalli-Sforza and Edwards, 1967) a parameter? (Felsenstein, 1978, 1988) or a model? (Kishino and Hasegawa, 1990b). Because topologies are discrete, numerical optimisation procedures described above cannot be applied. Secondly, the number of possible topologies increases quickly with the number of sampled taxa. There is only one unrooted tree with three taxa; the next taxon can be added on one of the three branches (three possibilities); there are five possibilities for the fifth branch; seven for the sixth, and so on. With s taxa, there are $1 \times 3 \times 5 \times \ldots \times (2s-5)$ bifurcating unrooted trees (Cavalli-Sforza and Edwards, 1967). This amounts to 105 trees with 6 taxa, 2, 027, 025 trees with 10 taxa, and 221, 643, 095, 476, 699, 771, 875 (about $2 \cdot 10^{20}$) with 20 taxa. A similar reasoning shows that this number increases to $\prod_{i=3}^{s}(2i-3)$ for bifurcating rooted trees. For instance, there are about $0.8 \cdot 10^{22}$ possible topologies with 20 taxa. Practical formulae can be found in Cavalli-Sforza and Edwards (1967); the gamma function can also be used, and the previous counts reduce to

$2^{s-2}\Gamma(s-3/2)/\sqrt{\pi}$ and $2^{s-1}\Gamma(s-1/2)/\sqrt{\pi}$, respectively. This representation has the merit of showing the rate of increase of the number of possible trees with the number of taxa considered, faster than exponential (recall Stirling's approximation: $\Gamma(s+1) \approx (s/e)^s\sqrt{2s\pi}$ for large s). As a result, even with modest species counts, an exhaustive enumeration can be tedious. Moreover, deriving the log-likelihood functions to compare them analytically becomes quickly complicated, even in a very simple case (Yang, 2000a). This motivates the use of heuristic methods, which are not specific to ML methods. Three general approaches have been described (Swofford et al., 1996): (i) stepwise addition, where to three initially selected taxa is added a new one in the position that maximises the likelihood score; (ii) star decomposition, which starts from a star tree progressively resolved as in (i); (iii) perturbation methods, where the tree is modified according to a branch-swapping algorithm. One such algorithm, used in Chapter IV, is the nearest-neighbour interchange (NNI) (see Figure I.2). Other algorithms exist, such as the subtree pruning regrafting (SPR) or the transversal bisection and reconnection (TBR) algorithms (Swofford et al., 1996, p. 485; Nei and Kumar, 2000, p.126). The choice of the algorithm is considered important (Swofford et al., 1996), as the simplest (NNI) has only two neighbouring trees for each internal branch, when TBR for instance has more. It is often advised to mix branch-swapping methods so that the whole tree space can be explored more efficiently (Swofford et al., 1996). While Takahashi and Nei (2000) have shown that NNI + SPR searches are as efficient as TBR, convergence to the best tree is still an issue, especially for algorithmic (i.e. non-probabilistic) methods. It is always good practice to perform several searches.

Most of these optimisation algorithms assume that the likelihood surface has a single mode. Multiple peaks can cause optimisation algorithms not to converge to the global

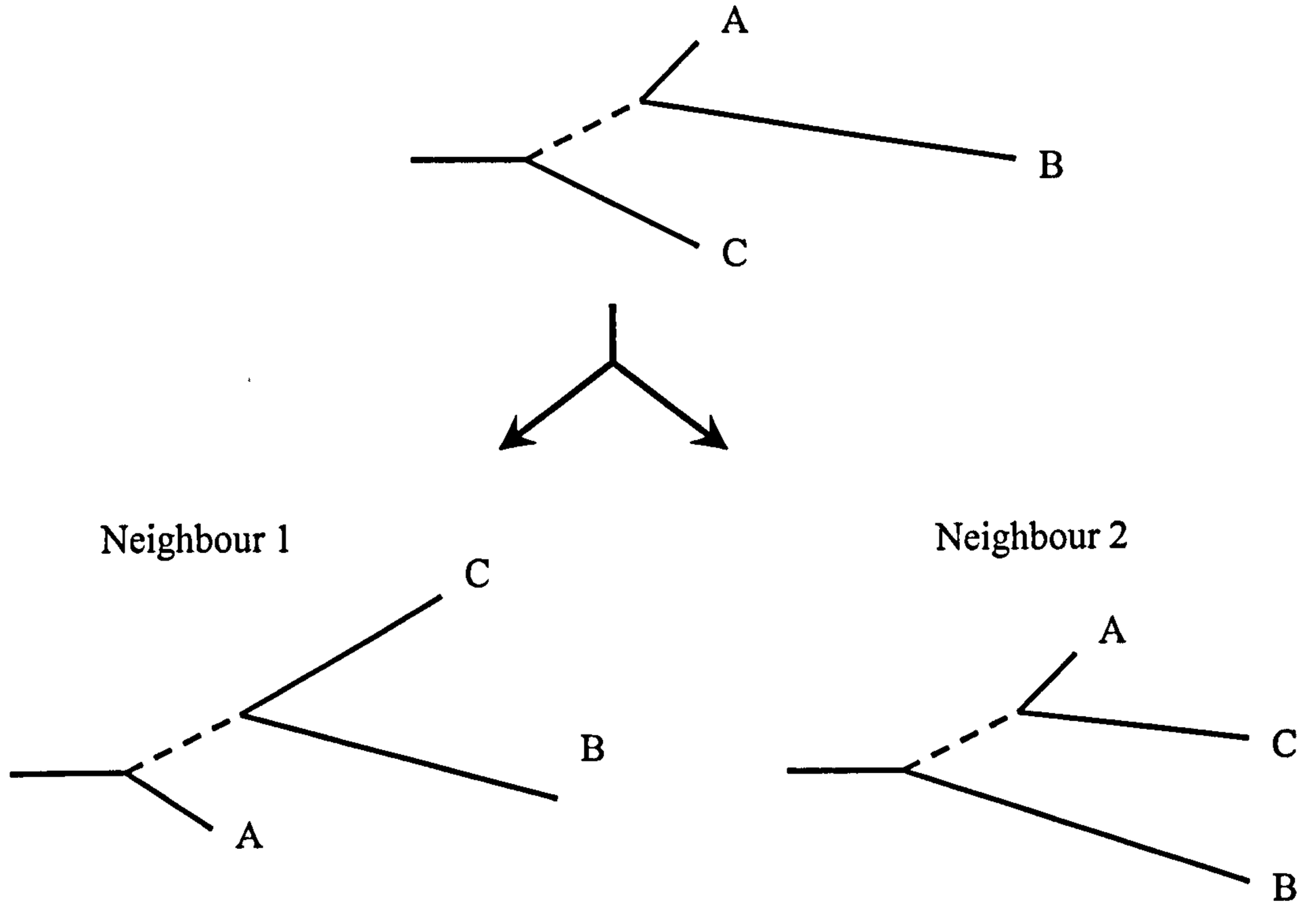


Figure I.2. Schematic representation of the NNI tree-perturbation algorithm: branches are swapped around the active branch (broken line).

peak (Steel, 1994). When topologies are quite "reasonable", i.e. for trees other than the ML tree but with high likelihood value, the likelihood surface appears relatively smooth, even when the substitution model is misspecified, but tends to become rugged for very unlikely trees, exhibiting multiple local optima (Rogers and Swofford, 1999). Note however that these authors never considered the possibility of local entrapment of the iterative algorithm on boundaries of the parameter space (branch lengths of zero) (Z. Yang, personal communication).

So far I have considered that the sequences are correctly aligned. Multiple sequences alignment can be difficult. This issue will not be dealt with in this thesis, but methods are being developed to integrate over some uncertainties of the procedure (see Lee (2001) for a review, and McGuire et al. (2001) for a model that integrates over iid gaps). Another assumption common to the models reviewed so far is the constancy of rates of evolution in space (among sites) and time (among lineages). The next section reviews how this assumption of rate constancy can be partially relaxed.

I.1.b – Markovian models of DNA sequence evolution

The likelihood of a tree is the probability of the data given a parametric model of evolution. With nucleotide data for instance, this model specifies the probability of a change from one nucleotide state to the next between any two points on the tree. This probability depends on two factors: the instantaneous rate of change from state *i* to *j* (denoted q_{ij}) and the amount of time (t) elapsed between these two points. The same data can either be explained by high rate and short duration, or vice-versa. This implies that branch lengths of a phylogenetic tree represent the expected number of substitutions per site, μ, with no implication with respect to the actual amount of time – unless the molecular clock can be assumed (see below I.2.a). To simplify the

interpretation of the results, rates are rescaled so that the average rate of change is one, that is:

$$-\sum_i \pi_i q_{ii} = 1 \qquad \text{(I.5)}$$

where π_i is the equilibrium frequency of nucleotide *i*.

A Markov process continuous and homogeneous in time (the transition mechanism does not change with time) is used to model the substitution process with a 4×4 rate matrix Q = $\{q_{ij}\}$. Its elements represent instantaneous substitution rates among the four nucleotides. Diagonal elements of the matrix are determined by the requirement that rows sum to zero, i.e. $q_{ii} = -\sum_{j \neq i} q_{ij}$ (e.g. Cox and Miller, 1965, p.180). The most general form of Q leads to the "unrestricted time-homogeneous model". Its rate matrix for nucleotide substitutions is given in Figure I.3, where nucleotides are ordered T, C, A, G and the q_{ij} coefficients are non-negative (Tavaré, 1986; Swofford et al., 1996, p.432). The unrestricted time-homogeneous model (Q_{UNREST} in Figure I.3) can be seen as a special case of the parameter-rich model of Barry and Hartigan (1987b), which assigns an unrestricted Q matrix to each branch of the tree. Their model is time-inhomogeneous. The distinction between the "frequency parameters" π_i's and the "rate parameters" *a* to *f* (see Figure I.3) may not represent distinct biological processes, but this generally makes the question mathematically more tractable (Yang, 1994a). Also mathematically convenient is the restriction that Q satisfies the reversibility criterion specified by the detailed balance equations:

$$\pi_i \, q_{ij} = \pi_j \, q_{ji} \qquad \text{(I.6)}$$

(e.g. Tavaré, 1986), and $\boldsymbol{\pi} = \{\pi_i\}^T$ is the stationary distribution of the process. This model, called the general reversible model (REV, or GTR), has now eight free parameters (five relative rates with *f* = 1 plus three base frequencies), instead of

$$Q_{JC69} = \begin{pmatrix} \cdot & 1 & 1 & 1 \\ 1 & \cdot & 1 & 1 \\ 1 & 1 & \cdot & 1 \\ 1 & 1 & 1 & \cdot \end{pmatrix} \qquad Q_{K80} = \begin{pmatrix} \cdot & \kappa & 1 & 1 \\ \kappa & \cdot & 1 & 1 \\ 1 & 1 & \cdot & \kappa \\ 1 & 1 & \kappa & \cdot \end{pmatrix}$$

$$Q_{F81} = \begin{pmatrix} \cdot & \pi_C & \pi_A & \pi_G \\ \pi_T & \cdot & \pi_A & \pi_G \\ \pi_T & \pi_C & \cdot & \pi_G \\ \pi_T & \pi_C & \pi_A & \cdot \end{pmatrix} \qquad Q_{F84} = \begin{pmatrix} \cdot & (1+\kappa/\pi_Y)\pi_C & \pi_A & \pi_G \\ (1+\kappa/\pi_Y)\pi_T & \cdot & \pi_A & \pi_G \\ \pi_T & \pi_C & \cdot & (1+\kappa/\pi_R)\pi_G \\ \pi_T & \pi_C & (1+\kappa/\pi_R)\pi_A & \cdot \end{pmatrix}$$

$\pi_Y = \pi_T + \pi_C$ (pyrimidines); $\pi_R = \pi_A + \pi_G$ (purines)

$$Q_{HKY85} = \begin{pmatrix} \cdot & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & \cdot & \pi_A & \pi_G \\ \pi_T & \pi_C & \cdot & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & \cdot \end{pmatrix} \qquad Q_{TN93} = \begin{pmatrix} \cdot & \kappa_1\pi_C & \pi_A & \pi_G \\ \kappa_1\pi_T & \cdot & \pi_A & \pi_G \\ \pi_T & \pi_C & \cdot & \kappa_2\pi_G \\ \pi_T & \pi_C & \kappa_2\pi_A & \cdot \end{pmatrix}$$

$$Q_{REV(GTR)} = \begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix} \qquad Q_{UNREST} = \begin{pmatrix} \cdot & q_{12} & q_{13} & q_{14} \\ q_{21} & \cdot & q_{23} & q_{24} \\ q_{31} & q_{32} & \cdot & q_{34} \\ q_{41} & q_{42} & q_{43} & \cdot \end{pmatrix}$$

Figure I.3. Rate matrices under different substitution models: Jukes Cantor (JC69), Kimura two-parameter (K80), Felsenstein (F81 and F84), Hasegawa, Kishino, Yano (HKY85), Tamura and Nei (TN93), general reversible model (REV, also noted GTR) and the unrestricted model (UNREST). Nucleotide states are ordered T, C, A and G. All the models are time-reversible (except UNREST) and nested from left to right and top to bottom: they are all special cases of UNREST. (Adapted from Yang, 1997b)

twelve parameters for the unrestricted model. The reversibility property is often assumed implicitly in comparing sequences from two contemporary species, since to get from one to the other one must travel backward in time to a common ancestor and then forward in time to the other species. This assumption makes the computation of the likelihood of a tree easier in that it does not depend on the position of the root (see above, section I.1.a).

While the general procedure to derive transition probabilities can be found in many textbooks (e.g. Cox and Miller, 1965), I give here a short outline that considers what happens to a nucleotide site. Following the general treatment of Markov processes, the probability of changing to state j at time t from state i at time s is noted $p_{ij}(s,t) = P\{X(t) = j \mid X(s)=i\}$. Between the two states i and j, the nucleotide site can be in any other state k at time u ($s < u < t$). The probability of this particular path is $p_{ik}(s,u)\, p_{kj}(u,t)$ since the two component probabilities are independent. The probability $p_{ij}(s,t)$ is obtained by considering all such possible paths, that is $\sum_k p_{ik}(s,u)\, p_{kj}(u,t)$. In matrix notation, this sum is written $P(t \mid s) = P(t \mid u)\, P(u \mid s)$, where $P(t \mid s) = \{p_{ij}(s,t)\}$. This is called the Chapman-Kolmogorov equation. It is used to derive a fundamental differential equation, given next, from which transition probabilities are obtained from rate matrices. By considering what happens between t and $t+\Delta t$ when leaving from s, $P(t + \Delta t \mid s)$ is $P(t + \Delta t \mid t)\, P(t \mid s)$. Let $\Lambda(t)\, \Delta t$ be a diagonal matrix where the probability of some change in the interval Δt out of state j is specified by the j^{th} element. Denoting $R(t)$ the limiting conditional transition matrix, $P(t + \Delta t \mid s)$ becomes $\{I - \Lambda(t)\, \Delta t\}\, P(t \mid s) + \{R(t)\, \Lambda(t)\, \Delta t\}\, P(t \mid s)$. Rearranging the equation and letting Δt tend to zero leads to the forward Kolmogorov equation: $\partial P(t \mid s) / \partial t = \{R(t) - I\}\, \Lambda(t)\, P(t \mid s)$. In the homogeneous case, $P(t \mid s) = P(t - s)$, and denoting $\{R(t) - I\}\, \Lambda(t)$ by $Q(t)$, the forward expression $dP(t) / dt = Q(t)\, P(t)$ is obtained (with the initial condition $P(0) = I$). If the matrix of infinitesimal transition probabilities,

Q, is time-homogeneous, the forward expression becomes $dP(t) / dt = Q\,P(t)$. A formal solution is:

$$P(t) = \exp(Q\,t) \qquad \text{(I.7)}$$

The computation of $\exp(Q\,t)$ usually follows a spectral decomposition that consists in obtaining the eigenvalues λ_i of Q to decompose this matrix as $Q = U\,D\,U^{-1}$, with $D = \text{diag}\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$. The matrix U contains the right eigenvectors x_i's, solution of the equations $Qx_i = \lambda_i\,x_i$ (i = 1 to 4). The λ_i's are the solutions of the determinantal equation $|Q - \lambda_i\,I| = 0$,which is a polynomial of degree four called the characteristic polynomial. $P(t)$ then becomes $U\,\exp(D\,t)\,U^{-1}$, i.e.:

$$U\,\text{diag}\{\exp(\lambda_i)\,t\}_{i=1,2,3,4}\,U^{-1}. \qquad \text{(I.8)}$$

Under some limiting conditions, it is also possible to approximate $P(t)$ by the series expansion:

$$\exp(Q\,t) \approx \sum_i (Q\,t)^i \qquad \text{(I.9)}$$

up to a certain degree (i). However, this approximation tends to be poor when branch lengths t are long.

Most of the substitution models described in the literature are special cases of the general reversible model (Figure I.3), and their corresponding transition probabilities are generally easy to obtain. They can be found in the literature (Tamura and Nei, 1993; Swofford et al., 1996, pp. 437-438), or by running the following script under Mathematica® for any Q matrix:

```
{evalues, evectors} = Eigensystem[Q];
b = Transpose[evectors];
b.(Exp[mu t DiagonalMatrix[evalues]] - {{0, 1, 1, 1}, {1,
0, 1, 1}, {1, 1, 0, 1}, {1, 1, 1, 0}}).Inverse[b];
Pmat = Simplify[%];
Pmat // MatrixForm
```

Because in Mathematica® the `Exp` function applied to a matrix $\{m_{ij}\}$ actually computes $\{\text{Exp}(m_{ij})\}$, the diagonal matrix $\exp(D\,t)$ is computed as `Exp[mu t`

`DiagonalMatrix[evalues]]` minus the matrix $\{1 - \delta_{ij}\}$, where $\delta_{ij} = 1$ if i =j, and $\delta_{ij} = 0$ otherwise (Kronecker function).

Calculating P(t) in the general reversible case is challenging, which motivates the use of a numerical algorithm (Yang, 1994a). However, noting that zero is the largest eigenvalue of Q (e.g. Cox and Miller, 1965, p. 183), finding the other roots reduces to solving a polynomial of degree three. The characteristic polynomial can then be written as $\lambda(A\,\lambda^3 + B\,\lambda^2 + C\,\lambda + D)$, where:

$$A = 1$$

$$B = f\,\pi_A + a(\pi_C + \pi_T) + b(\pi_A + \pi_T) + c(\pi_G + \pi_T) + d\,(\pi_A + \pi_C) + e(\pi_G + \pi_C) + \pi_G$$

$$C = a\,f\,\pi_A\,\pi_C + a\,e\,\pi_C^2 + c\,f\,\pi_A\,\pi_G + e\,f\,\pi_A\,\pi_G + a\,\pi_C\,\pi_G + e\,\pi_C\,\pi_G + a\,e\,\pi_C\,\pi_G + c\,e\,\pi_C\,\pi_G + c\,\pi_G^2 + e\,\pi_G^2 + c\,e\,\pi_G^2 + a\,f\,\pi_A\,\pi_T + a\,c\,\pi_C\,\pi_T + a\,e\,\pi_C\,\pi_T + a\,\pi_G\,\pi_T + c\,\pi_G\,\pi_T + a\,c\,\pi_G\,\pi_T + c\,e\,\pi_G\,\pi_T + a\,c\,\pi_T^2 + b\,(e\,\pi_A\,\pi_C + \pi_A\,\pi_G + e\,\pi_A\,\pi_G + a\,\pi_A\,\pi_T + c\,\pi_A\,\pi_T + a\,\pi_C\,\pi_T + e\,\pi_C\,\pi_T + c\,\pi_G\,\pi_T + e\,\pi_G\,\pi_T + a\,\pi_T^2 + c\,\pi_T^2 + f\,\pi_A\,(\pi_A + \pi_T) + d\,\pi_A\,(1 - \pi_G)) + d\,(e\,\pi_A\,\pi_C + e\,\pi_C^2 + f\,\pi_A\,(\pi_A + \pi_C) + \pi_A\,\pi_G + c\,\pi_A\,\pi_G + c\,\pi_C\,\pi_G + e\,\pi_C\,\pi_G + c\,\pi_A\,\pi_T + c\,\pi_C\,\pi_T + a\,\pi_C\,(1 - \pi_G))$$

$$D = a\,(\pi_G\,(f\,\pi_A + \pi_C + \pi_G + \pi_T)\,(e\,\pi_C + c\,\pi_T) + d\,\pi_C\,(\pi_A\,\pi_G + c\,\pi_A\,\pi_T + c\,\pi_C\,\pi_T + c\,\pi_G\,\pi_T + c\,\pi_T^2 + f\,\pi_A\,(\pi_A + \pi_C + \pi_T) + e\,\pi_C\,(\pi_A + \pi_C + \pi_G + \pi_T))) + c\,\pi_G\,(e\,\pi_G\,(f\,\pi_A + \pi_C + \pi_G + \pi_T) + d\,(f\,\pi_A\,(\pi_A + \pi_C) + \pi_A\,(\pi_G + \pi_T) + e\,\pi_C\,(\pi_A + \pi_C + \pi_G + \pi_T))) + b\,(d\,\pi_A\,(\pi_A\,\pi_G + c\,\pi_A\,\pi_T + c\,\pi_C\,\pi_T + c\,\pi_G\,\pi_T + c\,\pi_T^2 + f\,\pi_A\,(\pi_A + \pi_C + \pi_T) + e\,\pi_C\,(\pi_A + \pi_C + \pi_G + \pi_T)) + a\,\pi_T\,(f\,\pi_A\,(\pi_A + \pi_C + \pi_T) + c\,\pi_T\,(\pi_C + \pi_G + \pi_T) + \pi_A\,(\pi_G + c\,\pi_T)) + e\,(f\,\pi_A\,\pi_G\,(\pi_A + \pi_T) + (a\,\pi_C + c\,\pi_G)\,\pi_T\,(\pi_C + \pi_G + \pi_T) + \pi_A\,(\pi_C\,\pi_G + \pi_G^2 + a\,\pi_C\,\pi_T + c\,\pi_G\,\pi_T)))$$

In the most general case, the roots can then be calculated by means of a standard formula (run `Solve[x^3 + B x^2 + C x + D == 0, x]` under Mathematica®). One eigenvalue is real, while the two others are apparently conjugate complex numbers.

Direct computation of the eigenvalues of Q shows that, assuming that the q_{ij}'s and the π_i's are real, the complex roots of the above polynomial expression are of the form:

$$\alpha+(1+i3^{1/2})\,\beta/\gamma \pm (1-i3^{1/2})\,\delta/\varepsilon \qquad \text{(I.10)}$$

where the coefficients α to ε are non-linear cubic root functions of the polynomial coefficients. From here it is possible to check that imaginary parts are null, so that all the eigenvalues are real (see Yang, 1994a) and distinct. The eigenvectors and an algebraic expression of the transition matrix – although not simple – can be obtained for the general reversible model of nucleotide substitution, and should be checked against known results from simpler substitution models.

The substitution models presented so far are all time homogeneous. This assumption is known to be a very crude approximation (see Swofford et al., 1996), but very few alternatives have been proposed. Barry and Hartigan (1987a) proposed a parameter-rich model where each branch has a specific unrestricted rate matrix Q (with eleven parameters). A special case was proposed by Yang and Roberts (1995) where only different base frequencies are allowed for different branches under the HKY85 + Γ substitution model. More recently, Galtier et al. (1999) used a more specific a non-homogeneous and non-stationary Markov model developed earlier, allowing branch-specific CG contents (Galtier and Gouy, 1998; see also Lake, 1994 and Galtier and Gouy, 1995).

I.1.c – Modelling variable rates

Models described so far assume that sites are iid. In particular, it is assumed that all sites evolve at the same rate. This uniform mutation rate model has been challenged, as several studies suggested that the number of substitutions does not follow a Poisson distribution (Kocher and Wilson, 1991; Hasegawa et al., 1993; Wakeley, 1993). Instead, it appears to follow approximately a negative binomial distribution (Uzzell and Corbin, 1971), thereby

suggesting the use of a gamma distribution to model among-site rate variation (Johnson and Kotz, 1973; see Yang (1996a) for a review). If the probability of the data at site *i* given rate *r* is $p(x_i \mid r)$, its average over variable rates is:

$$\ell(x_i) = E_r[p(x_i)] = \int_0^\infty f(r \mid \alpha)\ p(x_i \mid r)\ dr \qquad \text{(I.11)}$$

where $f(r|\alpha)$ is the density of the gamma distribution with shape parameter α. Because *r* is a relative rate, the distribution chosen to model rate variation can be chosen to have mean one: the inverse scale parameter is set to α, and the variance is then $1 / \alpha$. The gamma distribution now has only one parameter, denoted α, which influences the shape of the distribution (e.g. Yang, 1993). Remarking that parameter estimates, such as the branch lengths, might be sensitive to the distribution used to model among-site rate heterogeneity, Kelly and Rice (1996) proposed a general model without assuming a parametric distribution. Their model is however heavily parameterised: each site is assigned a specific rate (or relative rate as the authors put the constraint that the mean rate is one). This may lead to unidentifiablility as there are more parameters than data points. A more detailed analysis might be needed to determine which model fits the data best, but it is likely that simple models such as the gamma distribution explain a substantial part of the observed variance with minimum bias. In the discrete version of the gamma model (Yang, 1994b), a finite number *n* of rate categories is defined. The rate of a particular site is averaged over these categories, and the likelihood of that site is:

$$\ell(x_i) = \sum_{j=1}^{n} p_j\ p(x_i \mid r_j) \qquad \text{(I.12)}$$

Each rate category r_j is assigned an equal probability $p_j = 1/n$ by using appropriate binning of the original (continuous) distribution. For example if $n = 8$, the set of r_j is the 1/16, 3/16, ..., 15/16 percentiles of the gamma distribution $f(r|\alpha)$ (Felsenstein, 2001). Note that partitioning (see below) and modelling among-site rate variation by means of a

gamma distribution are not exclusive. A similar approach has been applied to detect positive selection by modelling the ratio of the nonsynonymous to synonymous substitution rate (ω), assumed to be drawn from some distribution or some mixture of distributions (Nielsen and Yang, 1998). Again, it is possible to replace this approach with partitioning the data, each "gene" corresponding here to a distinct functional region with its own nonsynonymous to synonymous substitution rate (Yang and Swanson, 2002).

In protein coding sequences, some sites are expected to be under strong negative selection. They then appear invariant. Models accounting for this category have been developed on the basis of HKY85 and TN93 (Gu et al., 1995), usually denoted by the suffix "I" when referring to the model. The most general model (REV + Γ + I) has been described and used (e.g. Posada and Crandall, 2001a). It has been shown to have interesting properties (see below and Rogers, 2001).

Another possibility to deal with among-site rate variation is to partition the data, dividing them into several blocks such as first, second and third codon positions. For each partition, the general model uses parameters either shared or independent across partitions. The sole assumption is that branch lengths are proportional among the partitions (Yang, 1996b). This can be used in conjunction with the gamma distribution modelling among-site rate variation within a partition.

Over long periods of time, sites critical to protein function may change, hereby changing its mode (rate) of evolution. Note this is a discrete change, not a gradual one. This notion was called the "covarion" model (Fitch and Markowitz, 1970; Fitch, 1971), which stands for "*co*ncomitantly *vari*able cod*ons*". While the issue of among-lineage rate variation has been raised a long time ago (see above), covarion models received a closer scrutiny only recently. Recently, a covarion model has been proposed; based on hidden

Markov processes, only two parameters are added to the substitution model (Galtier, 2001; Penny et al., 2001).

I.2 – Estimating divergence dates: molecular clock and local clocks

One of the objectives of phylogenetics is to date particular events (e.g. Kishino and Hasegawa, 1990, p. 550). Here, I review methods commonly used to estimates divergence dates in a likelihood framework. I underline the importance of testing the assumptions made to derive the estimates.

I.2.a – The molecular clock hypothesis

A characteristic of ML-based models of evolution is that times and rates are indistinguishable, unless some specific assumption is made about the one or the other. Studies of the evolution of proteins such as hemoglobin, cytochrome C or fibrinopeptides suggested that they evolve with an *approximately* constant rate (Zuckerkandl and Pauling, 1962, 1965; Margoliash, 1963; Doolittle and Blombäck, 1964). Let $B(T)$ be the branch length, i.e. the number of substitutions occurring on a branch from time $t = 0$ to time T. If evolutionary events occur independently with rate $R(t)$ at time t, $B(T)$ follows a Poisson distribution with mean $\int_0^T R(t)\,dt$ (Gillespie, 1991; Thorne et al., 1998). If the rate of evolution is constant, branch lengths are proportional to time so that:

$$B(T) = \int R(t)\,\mathrm{d}t = R\,T \qquad \text{(I.13)}$$

If at least one of the divergence time in a constant-rate tree is known, by the existence of a precisely dated fossil, then it is possible to calibrate the clock and to estimate all the times of all divergences in the tree.

History of the molecular clock assumption

The molecular clock assumption was formulated when the synthetic theory of evolution was at its height, i.e. when the rate of evolution was supposed to be under the control of

environmental changes and natural selection. Morphological evidence was known to indicate variable modes and tempos of evolution (Simpson, 1944), so that the molecular clock has been controversial since the very beginning (Simpson, 1964; see Nei and Kumar, 2000). To explain the discrepancy between morphology and molecular-based studies, Kimura (1968, 1969) and King and Jukes (1969) proposed a hypothesis: if most of the amino acid substitutions are the result of random fixation of selectively neutral or nearly neutral mutations, then a uniform rate is expected. The implication is that the rate of neutral evolution is constant *per year* for a given protein. This has been found to range from $0.010 \cdot 10^{-9}$ substitutions per amino acid site per year for histone H4, to $9.0 \cdot 10^{-9}$ for fibrinopeptides (Nei, 1987). Kimura (1969) coined the term *pauling* to express this unit (10^{-9} substitutions per amino acid site per year) as an analogue to the *darwin*, the unit of evolutionary rate at the phenotypic level, representing the variation of a quantitative measurement at the rate of 10^{-3} per 10^{3} years (see Kimura, 1969).

Definition of the molecular clock

Substitutions do not occur at regular time intervals: the clock is erratic (Kimura, 1983). In its exact formulation, the clock assumption stipulates that the expected rate of substitutions B accumulated along a lineage i of length T is constant, that is:

$$\mathrm{E}[B_i(T)] = \lambda_i T \qquad \text{(I.14)}$$

with $\lambda_i > 0$ and $\lambda_i = \lambda_j$ for $i \neq j$. To demonstrate the validity of this assumption, one has to test whether the expected rates are equal among lineages. If it is assumed that substitutions are rare independent events occurring with a non-zero probability after a sufficiently long waiting time, then $B_i(t)$ naturally follows a Poisson distribution (Kimura, 1983). It is this supplementary assumption that lead Ohta and Kimura (1971) to focus on the ratio I of the observed to the expected variance under the Poisson process as a means to test the molecular clock assumption. As $B_i(t)$ follows a Poisson distribution, the

expected variance is equal to the mean, and if n denotes the number of lineages, $(n-1)\,I$ is distributed as a χ^2_{n-1} (Kimura, 1983) under the assumption of a star tree (Goldman, 1994). This ratio I is called the dispersion index of the stochastic process considered (Poisson here). Its use is facilitated by assuming that the number of substitutions accumulated between two lineages, i and j, equals $B_i(t) + B_j(t)$ (Gillespie, 1991; Zheng, 2001). Kimura (1983) then operated a conceptual shift from *hypothesis testing* to *estimating* the dispersion index. This lead to a long lasting debate (i) on the way to estimate I (see Gillespie, 1991), (ii) on the distribution followed by different estimators of I when a non-star phylogeny is considered (Goldman, 1994), (iii) on the overdispersion of the molecular clock (Takahata, 1987; Gillespie, 1988; Takahata, 1991a,b) and (iv) on possible means to model departures from the clock by modifying the original Poisson process. Such departures can be modelled with a Cox process (Gillespie, 1991) or related models (Cutler, 2000a,b,c).

Dates estimation under the molecular clock

Estimating divergence dates under the clock raises other issues. Besides the problem of the exact position of the calibration point and its accuracy (Springer, 1995; Lee, 1999), characterising the timescale of evolution of any group of taxa demands good estimates of rates of evolution, at least on a relative scale (Gillespie, 1991). However, these estimates are often obtained by pairwise comparisons, which average rates over the entire phylogeny (Ayala et al., 1998). The most popular method consists in estimating the slope of the regression of pairwise distances against calibrated divergence time (see Martin, 2001). Besides the problems of phylogenetic non-independence (Pagel, 1997; Ayala et al., 1998; Sanderson, 2002) and of the loss of information due to pairwise comparisons, the node of interest is often the root, whose estimated divergence is often at a worrying distance of the data points (e.g. see results from Wray et al., 1996; Leitner and Albert,

1999; Korber et al., 2000), making the estimate highly unreliable (e.g. Sokal and Rolf, 1995).

While the molecular clock hypothesis seems to perform well in some cases (Kimura, 1983; Nei and Kumar, 2000), analyses of most data sets show this assumption is often violated. Showing this requires powerful tests of the clock, presently presented.

I.2.b – Testing the molecular clock assumption

Independently of the dispersion index, a variety of statistical tests were proposed to evaluate the molecular clock hypothesis. Most have an intuitive basis (Sarich and Wilson, 1967), latter formulated in statistical terms (Langley and Fitch, 1974; Fitch, 1976). The most recent ones belong to four classes: distance-based tests, also called relative rate tests (RRT), least-square methods, two-cluster and branch-length tests, and likelihood ratio tests (LRT).

Three-species methods or relative rate tests (RRT)

The RRT works on pairs of taxa and compares the distance from a member of the pair to an outgroup taxon. For this reason, this type of test is also called "triplet test", or three-species test. The approach can be used with different distances, can make use of different test statistics, and can be performed on different types of sites. Most of the RRT (Wu and Li, 1985; Tajima, 1993a; Li and Bousquet, 1992) have been described with Kimura's two-parameter distance (see section I.1.b; Kimura, 1980), for which variances and covariances are readily available (e.g. Nei and Kumar, 2000). Standardised differences are used to compute the significance level by Wu and Li (1985): if we assume that substitutions follow a Poisson process, the significance level is determined by a normal test (Z test). Tajima (1993a) proposed two chi-square tests which do not require the variance of the distances. One tests the equality of the expected distances computed for

all "informative" (see below) sites (1D method, as there is one degree of freedom), while the other partitions sites between transitional and transversional differences (2D method, which has two degrees of freedom).

Different methods consider different classes of sites. In the version by Wu and Li (1985), distances are computed for non-synonymous and synonymous sites. These two classes are subdivided into non-degenerate, two-fold and four-fold degenerate sites. Non-degenerate sites are those for which any substitution is either non-synonymous or nonsense. At two-fold degenerate sites, a change is synonymous if it is a transition (C↔T or A↔G) and is non-synonymous (or nonsense) if it is a translation. The two exceptions are (i) in the nuclear genetic code, the first position of four of the six arginine codons and (ii) the third position of the three isoleucine codons are three-fold degenerate. Adjustments have been made to accommodate these peculiarities (Li et al., 1985). Tajima (1993a) considers only "informative sites", defined as those in which the outgroup and one of the ingroup taxa share a character state, the other ingroup taxon being in a different state. This test ignores sites where three different states occur, as this configuration is not informative for testing an excess of substitutions in one of the ingroup taxa. Therefore, this test has its maximum power for binary characters; four-state (nucleotides), twenty-state (amino acids) and sixty-one-state characters have reduced power because the probability of observing three different states increases, thereby reducing the number of sites considered (Bromham et al., 2000). This holds for both methods, 1D as well as 2D, even if the latter is less conservative (Tajima, 1993a). It is important to understand the limitation of such tests and the extent of their lack of power, which may be confounded with the issue of multiple testing. In particular, Bromham et al. (2000) showed that the RRT (Tajima, 1993a) is unlikely to detect among-lineage rate variation up to a factor four, which may seriously bias estimates of divergence dates.

Moreover, the reliability of triplet tests depends on the taxa and the outgroup chosen, so that they do not always provide a suitable filter for removing rate-variable sequences from the analysis (Bromham et al., 2000). This points to the importance of using more powerful tests, which also avoid the issue of multiple tests.

Least square methods

A different approach, which will be extended below in the likelihood framework, is to compare the fit of the model to the data with and without the clock assumption. When branch lengths are estimated with the method of least squares (e.g. Li, 1997, p.125; a computational method is given in Rzhetsky and Nei, 1993; see Uyenoyama, 1995 for an application of the generalized least square method), it is possible to test the clock by comparing the least square residual sums under the assumption of rate constancy (R_C) or without (R_N) (Felsenstein, 1984;Felsenstein, 1988). The statistic:

$$\{(R_C - R_N) / (s - 2)\} / \{R_N / [s\,(s - 1)/2 - (2s - 3)]\} \qquad \text{(I.15)}$$

is assumed to follow a F distribution with the degrees of freedom $s - 2$ and $s\,(s - 1)/2 - (2s - 3)$, where $s \geq 4$ species. It is implicitly assumed that pairwise distance estimates are iid and follow a normal distribution. While normality holds when the number of substitutions is large, pairwise distances are positively correlated because of the underlying phylogeny. Consequently, the reliability of this test is unclear (Felsenstein, 1988, 1995).

Two-cluster and branch-length tests and linearized trees

Takezaki et al. (1995) proposed two tests. The three-species tests described above can be seen as special cases of the two-cluster test, where each group (cluster) of taxa contains a single sequence. For each cluster, the quantities compared are the average of the

estimated distances. As with the method by Wu and Li (1985), it is necessary to have an expression of the variance-covariance of the distance chosen in order to compute the significance level (Z test). The branch-length test differs from the two-cluster test in that it evaluates the difference between the root-to-tip sum of branch lengths and an average for all the sequences but the outgroup. Branch lengths are estimated by the ordinary least-square method, which gives an estimate of the standard errors (Rzhetsky and Nei, 1993). Standard errors can also be derived by bootstrap (Takezaki et al., 1995). This difference is tested by means of a Z test. These two tests are sometimes referred to as "phylogenetic tests" (Nei and Kumar, 2000, p.196).

The tests described above are usually used to remove taxa exhibiting variable rates from a data set. This is achieved by performing a series of RRT from the tips of the tree to the root (the issue of multiple comparisons is seldom raised, though). Clusters showing a significant rate difference are iteratively removed from the data set. This results in a linearized tree, on which a clock-like analysis can be performed (Takezaki et al., 1995). The price of such a procedure is to lose a number of taxa from the analysis.

Maximum likelihood

This method, originally described by Felsenstein (1981, 1983, 1988; see also Kelly and Rice, 1996), requires the maximisation of the likelihood of a set of sequences under a given (unrooted) tree topology, with and without the constraints implied by the clock assumption. These two models are nested, so that the likelihood ratio test (LRT) statistic (the deviance, or minus twice the log-likelihood difference) follows asymptotically a chi-square (χ^2) distribution when the clock model is correct (e.g. Shao, 1999, p. 384; see below section I.3.a). In the case of reversible models (section I.1.b), with *s* species, there

are $2s - 3$ branch lengths when the clock is not assumed, and only $s - 1$ node times under the clock. The degrees of freedom are then $s - 2$.

Muse and Weir (1992) also presented a method using a LRT to test the clock, but following the principle of "three-species" tests. The idea is to estimate branch lengths with and without the clock for a triplet of taxa in a likelihood framework. The equality of the branch lengths leading to the two ingroup taxa is tested by a LRT, whose test statistic is assumed to follow a χ^2 distribution. As above, there are $s - 2$ degrees of freedom, which reduce to one in this case ($s = 3$). The Muse and Weir (1992) test is more flexible than the RRT, based on variances-covariance matrices, as it allows substitution models where these quantities are difficult to obtain. This three-species LRT is also more powerful than the RRT tests (Muse and Weir, 1992), although its power relative to the LRT on the whole data set has not been investigated.

The use of the χ^2 as the limiting distribution of the deviance was questioned (Goldman, 1993). Firstly, the number of site patterns increases rapidly with the number of sequences (4^s). The rule of the thumb that requires that each pattern be observed a few times (five in general) is then rarely satisfied. A solution is to resort to parametric bootstrap to estimate the distribution of the test statistic under the null hypothesis (Goldman, 1993). Assuming the model is correct (topology and substitution model), the branch lengths estimated for the original data set under the clock (null hypothesis) are used to simulate a large number of replicates. Simulated data sets are then analysed with and without the clock, in order to obtain the distribution of the deviance under the null hypothesis. This test is more specific than the LRT of the clock (Felsenstein, 1981) as rejection of the null hypothesis suggests that one or more of its components are inadequate. While the computational cost of this method is not negligible, Goldman (1993) has noted that the χ^2 approximation to the distribution of the deviance is

reasonable. Yang et al. (1995) suggest that this approximation is not severely affected by our lack of knowledge about the correct model. Secondly, there are some problems associated with the tree topology, assumed to be true in the LRT of the clock (see Goldman, 1993; Yang et al., 1995). Using a wrong tree may give overconfidence in a result (Yang et al., 1994), which may be the "wrong" one (see Chapter IV). A more general problem with the LRT is its sensitivity to the model, in particular to the tree topology (Yang et al., 1995).

The failure of a global molecular clock is usually explained by four hypotheses: the effects of generation time, metabolism, DNA repair and natural selection. These hypotheses are not mutually exclusive and are probably not exhaustive. Briefly, when DNA is replicated, the polymerase introduces some copy errors. If we assume that errors are generated at the same rate across taxa, those with a larger number of germ-line cell divisions accumulate more copy errors. This is the origin of the male-driven evolution theory (Miyata et al., 1987). The clock should run faster in organisms with short generation times as their germ-line cells undergo more rounds of duplications. This generation time hypothesis has been used to explain why hominids seem to be evolving more slowly than rodents at the molecular level (Kohne, 1970; Li et al., 1996). The metabolic rate hypothesis proposes that higher metabolic rates cause higher levels of oxidative damage to DNA, ultimately leading to different substitution rates (Martin et al., 1992; Martin and Palumbi, 1993; Rand, 1994). Others have proposed that the efficiency of the repair mechanism varies across taxa (Drake and Baltz, 1976; Goodman et al., 1984; Britten, 1986). Varying selective pressures during evolutionary time also affect the clock. While changes at non-synonymous sites (see section V.3) are expected for instance after duplication events (Force et al., 1999; Lynch and Force, 2000), synonymous sites

may also be subject to purifying selection (Hurst and Pal, 2001 but see Urrutia and Hurst, 2001) because of codon bias (Marais and Duret, 2001) or biophysical constraints on chromosomes (see Martin, 2001).

I.2.c – Local molecular clocks

In front of the overwhelming evidence that there is no universal clock, alternative models of sequence evolution must be developed. The hypotheses explaining rate variation suggest that closely related species are likely to evolve approximately at the same rate, so that the global tree would have local molecular clocks. This idea was first put into practice to date the divergence dates among primates (Hasegawa et al., 1989), as there is evidence of a rate slowdown in the Hominoidea when compared to Cercopithecoidea (e.g. Li and Tanimura, 1987). The idea underlying the approach of Hasegawa et al. (1987, 1989) is to reduce the data to transition and transversion differences between pairs of species. The vector $\boldsymbol{D}$ of transitions and transversions as well as the variance-covariance of $\boldsymbol{D}$ are computed under the HKY85 substitution model (Hasegawa et al., 1985). Local clocks are allowed by setting different transition rates (α) and transversion rates (β) to different sets of branches. For instance, one of the models described by Hasegawa et al. (1989) assumes (α_1,β_1) for the Hominoidea and (α_2,β_2) for the Cercopithecoidea. A multivariate normal approximation is then used to approximate the distribution of $\boldsymbol{D}$ for all the species pairs. Parameters are estimated iteratively by the generalised least-squares method. An ML version of the model with the same approximation was subsequently proposed (Kishino and Hasegawa, 1990b) and applied to estimating the timescale of human mitochondrial evolution (Hasegawa et al., 1993). This approach requires that the tree topology is known, but this is common to most of the methods (but see Chapter VI). However, a drawback of the method is that branch length estimates are based on pairwise comparisons, with a consequent loss of information (Penny, 1982).

Based on a method presented by Cooper and Penny (1997), Rambaut and Bromham (1998) developed a somewhat more general likelihood model to estimate divergence times (see also Bromham and Hendy, 2000). The idea is to consider quartets of taxa forming rooted and symmetrical trees. Each pair of taxa is assumed to evolve at a rate constant down to the root, and must have a known fossil-derived divergence date. As there are only two rates on the quartet, this model is referred to as the "two-rate model" (Rambaut and Bromham, 1998). The fit of the model can be assessed by a likelihood ratio test. The date of the root and the two rates are estimated by maximising the likelihood of the quartet. Confidence intervals are readily obtained. In a study with more than four taxa, this basic operation is repeated by combining into quartets all the pairs of taxa with fossil-derived divergence dates. Quartets where the "two-rate model" is rejected are discarded. Simulations suggest that the method is robust to moderate departures from both rate heterogeneity and substitution process (Rambaut and Bromham, 1998). This model was applied by the same authors to estimate the origin of the modern birds orders, and the origin of the Metazoa (Bromham et al., 1998; see Chapters II and III). However, this model has some limitations. For instance, it is well adapted to estimating the divergence of the origin of a group of taxa, but because of its quartet structure, a large amount of taxa with known fossil-derived divergence dates must be incorporated in the analysis. Moreover, a substantial amount of information may have to be discarded when testing the clock. It is also possible that the use of quartets leads to some other complications, not yet well understood (e.g. multiple testing).

To overcome such difficulties, Yoder and Yang (2000) have implemented a ML model of local molecular clocks on the entire tree. This model assumes that some branches, e.g. a group of related species, evolve at the same rate, while other sets of branches evolve at different rates. Let us choose, a priori, *k* sets of branches. One such set

is defined to evolve at a basal relative rate $r_1 = 1$, while the other $k-1$ rates are multiplication factors. For all the local rate parameters to be identifiable, the total number of parameters in a local clock model cannot exceed that of the unconstrained model: with s species under local clocks, there are $s-1$ node times and the total number of parameters is then $s+k-2$, whereas an unrooted tree has $2s-3$ branch lengths. The maximum number of local rates is therefore constrained by: $s+k-2 \leq 2s-3$., that is, k cannot exceed $s-1$ (or a basal relative rate $r_1 = 1$ and $s-2$ multipliers). This model was used to estimate the dates of the rat-mouse divergence and of some primates from the mtDNA of 31 mammals. While estimates for the primates were consistent with the fossil record, the rat-mouse estimates appeared to be problematic (Yoder and Yang, 2000; compare with Appendix 1). A possible reason may be that local clocks do not encompass all the rate variability of the data. Moreover, branches of the tree evolving at different rates are chosen on an a priori basis, which can be a pre-analysis of the same data set. While the number of sets of branches with pre-assigned rates is large (Sanderson, 1998, 2002), the statistical limitation of this approach is unclear.

I.3 – Testing further evolutionary hypotheses

In any statistical analysis, it is traditional to distinguish between random and systematic errors (Swofford et al., 1996, pp. 493-509). By "error" it is usually meant "*deviation between a population parameter and the true value of that parameter*" (Swofford et al., 1996, p. 493), that is the difference between the estimation of a real-valued parameter θ of an unknown population and θ. Random errors are due to our finite sampling. If a consistent method is used, these errors should vanish as sequence length increases and tends to infinity. On the other hand, when the model is misspecified, the "true" parameter cannot be estimated, which generates systematic errors. Increasing the sample size in this

case does not reduce systematic errors. The consistency of ML methods (Wald, 1949) may prevent "positively misleading" inference (Felsenstein, 1978), but wrong models will be shown to lead to overconfidence (see Chapter IV). Still, this is probably preferable to using non-statistical methods such as parsimony (Swofford et al., 2001; Sullivan and Swofford, 2001), as taking some complexities into account, such as a general reversible model (see section I.1.b) accommodating for among-site rate variation (I + Γ; section I.1.c) can lead to fully consistent results (Rogers, 2001).

Evolutionary hypotheses include many questions: some are easy to assess, while others are more delicate. Testing the molecular clock for instance (see above, I.2.b) belongs to the first category, as the model describing the clock assumption is nested in the more general model where rates are free to vary. More difficult questions will be addressed in the following section, together with the general testing approaches used so far. In particular, one of the most difficult question concerns the topology itself (Yang et al., 1995). This issue will be dealt with in more detail later (see Chapter IV).

I.3.a – Comparing nested models

Let us consider first the case of testing simple hypotheses, that is hypotheses which are completely specified by the distribution of the random variable of interest. Two types of errors are possible: either reject the null hypothesis H_0 when is it true (type I error), or fail to reject it when it is false (type II error). These errors occur with probabilities α and β, respectively. When testing a simple null hypothesis H_0 against a simple alternative H_1, it is not possible to keep both α and β arbitrarily small: decreasing one increases the other. The option is usually to fix α to a small value, and attempt to control β by choosing a test that maximises the power $1 - \beta$. An important result of statistical theory is the Neyman-Pearson lemma (e.g. Lehmann, 1959, p. 63). In the continuous case, let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a vector of n iid variables with an unknown density function $f_X(\cdot)$. The null hypothesis

H_0 specifies that **X** follows $f_0(\cdot)$, whereas the alternative specifies it follows $f_1(\cdot)$. The Neyman-Pearson lemma states that the most powerful test of a null hypothesis H_0 against H_1, that is the test that for a given fixed type I error α maximises the power $1 - \beta$, is obtained by using the likelihood ratio as a test statistic:

$$LR(\mathbf{X}) = f_0(\mathbf{X}) / f_1(\mathbf{X}) \qquad \text{(I.16)}$$

This is easily extended to composite hypotheses (i.e. hypotheses that are not simple) with the generalised likelihood ratio test statistic:

$$GLR(\mathbf{X}) = \max_{\omega}[f_0(\mathbf{X})] / \max_{\Omega}[f_1(\mathbf{X})] \qquad \text{(I.17)}$$

where each density (likelihood) is maximised on a parameter space, H_1 being characterised on Ω and H_0 on ω. Note however that the optimality of the Neyman-Pearson lemma no longer holds in the case of composite hypotheses (e.g. Lehmann, 1959, p. 63). In the nested case, ω is a subspace of Ω, or equivalently, H_0 is a particular case of a more general hypothesis H_1. The aim of hypothesis testing in the nested case is somewhat different than in the non-nested case: in the nested case, we are interested in knowing whether the more general model explains the data significantly better than the simpler model. Besides this specific objective, the reason to focus on nested models is that the deviance $-2\,LR(\mathbf{X})$ (see above, I.2.b) is asymptotically distributed as a χ^2. The degree of freedom is the difference between the number of parameters specified by the null and by the alternative hypotheses (Wilks, 1938). This result is derived under some regularity conditions, crucial to the correct application of the approximation. An intuitive outline is provided here. Let θ be a parameter, either a scalar or a vector, that takes a given value θ_0 under H_0 and is left unspecified under H_1, where the MLE is θ^*. The deviance is $2(\ell(\theta^*; \mathbf{X}) - \ell(\theta_0; \mathbf{X}))$, where ℓ is the log-likelihood function. Under H_0, $\theta = \theta_0$, and for large n, the Taylor expansion of $\ell(\theta_0; \mathbf{X})$ about θ_0, together with the fact that the derivative of ℓ is zero at the MLE, makes the deviance approximately equal to

$-(\theta^* - \theta_0)^2 \, d^2\ell / d\theta^2$. The expectation of $-(d^2\ell / d\theta^2)^{-1}$ is the variance of θ, so that the deviance is approximately $(\theta^* - \theta_0)^2 / \text{var}(\theta)$, which is approximately distributed as a χ^2 with one degree of freedom. This completes the simplified version of the proof. Two important requirements, part of the regularity conditions for this approximation to hold, are (i) that the parameters involved in the test are real-valued, and (ii) that none of them is on the boundary of the parameter space, a space where the likelihood function is otherwise differentiable.

This χ^2 approximation is mainly used to test the molecular clock hypothesis ("general adequacy test": see above I.2.b), or whether a substitution model explains the data significantly better than a simpler model (Huelsenbeck and Crandall, 1997) ("comparision of two parametric models"). This is valid because (i) the substitution models are nested (they all appear as special cases of the REV + Γ + I – e.g. Posada and Crandall, 2001a; see I.1.b) and (ii) the tree topology, whose nature is discrete, is fixed (Huelsenbeck and Rannala, 1997). Whelan and Goldman (1999) give a simulation example to show that when the two models are not nested, the χ^2 approximation does not hold – which was not unexpected (e.g. Vuong, 1989). "Nestedness" is one of the crucial regularity conditions for the χ^2 limiting distribution to hold (see above). Other issues have been raised about the validity of this approximation (Goldman, 1993; see I.2.b), but real data analyses (Yang et al., 1995) as well as simulation studies (Whelan and Goldman, 1999) suggest the approximation is reasonably good. Testing the fit of substitution models still receives some careful attention (Posada and Crandall, 1998, 2001a,b). Note however that when the simpler model is not rejected, the implication is that the more sophisticated model does not provide a significantly better fit to the data, and not necessarily that the simpler model gives a satisfactory explanation of the data. In the case of nested models, the LRT is a comparative test (e.g. Ewens and Grant, 2001, p.418).

Another requirement of Wilks' limiting distribution is that parameters must not be at any boundary, a situation that arises e.g. when testing the fit of models incorporating among-site rate variation. The gamma distribution (Yang, 1994b) usually improves dramatically the likelihood value of the model, but neglecting among-site rate variation is equivalent to assuming that the shape parameter α is infinity, which is a boundary of the real line. Whelan and Goldman (1999) showed, using simulations, that the deviance does not follow a χ^2 distribution with the expected degree of freedom. These authors claimed that the time consuming parametric bootstrap is the only rigorous way to test the fit of a model incorporating among-site rate variation, when the approximation is assumed to be valid (Yang, 1997a). Self and Liang's (1987) results show that in testing the fit to the gamma distribution, the deviance follows a mixture of χ^2 distributions, $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. This result was confirmed by simulations (Goldman and Whelan, 2000), but holds when only one parameter is at the boundary. When more parameters are at the boundary of the space, the deviance generally follows a mixture of χ^2 distributions whose components are not easy to calculate (Self and Liang, 1987; Ota et al., 2000). This is all the more problematic as issues involving models with two or more parameters at the boundary are appearing (Huelsenbeck and Nielsen, 1999; Yang et al., 2000).

I.3.b – Comparing non-nested models

When models are nested, testing which model significantly fit the data best is not complicated. However, when models are not nested, the deviance is not χ^2 distributed. One of the cases in which this is relevant to phylogenetic issues is in testing topological differences. Although here most methods and tests are derived from Bayesian arguments (Westfall and Young, 1993, p. 22), I will separate classical ML treatment from information-theoretic ones. Within the classical ML treatment, a distinction between parametric and non-parametric methods is made.

Maximum likelihood approach

Maximum likelihood is now widely considered a justified statistical method (Felsenstein, 1983), even if some have questioned its validity when applied to phylogenetic questions. In particular, as the likelihood function changes from topology to topology, Nei (1987), Saitou (1988) and Nei and Kumar (2000) see the maximum likelihood values for different topologies as conditional probabilities that cannot be compared in a usual statistical sense. However, it is possible to define a measure between the true unknown distribution $f(\cdot)$ and an approximating model $g(\cdot \mid \theta)$, with the following expectation (with respect to $f(\cdot)$):

$$I(f,g) = \mathrm{E}_{f(\cdot)}\{\log(f(X) / g(X \mid \theta))\} \qquad \text{(I.18)}$$

This measure, called the Kullback-Leibler information, or K-L distance (Kullback and Leibler, 1951), is non-negative, which makes it a good candidate to compare possibly non-nested models (Akaike, 1973; Kishino and Hasegawa, 1989). Note this distance is not a metric as it does not satisfy the triangular inequality and $I(f,g) \neq I(g,f)$. A simple approximation is derived as the likelihood value at the MLE, $\ell(\theta^* \mid X)$ (see below). Observing that maximum likelihood is dependent on the realisation of a particular tree, it is possible to define and to estimate the variance of $\ell(\theta^* \mid X)$ of different trees (Kishino and Hasegawa, 1989). This variance was originally obtained following a Bayesian argument: Kishino and Hasegawa (1989) take the posterior probability of a given tree T_i as $\mathrm{p}(\theta_i \mid X)\ \mathrm{p}(T_i) / \mathrm{p}(X)$, where $\mathrm{p}(X)$ is a sum over the tree space (see Chapter IV and equation IV.5). This is estimated as:

$$\exp(\ell_i(\theta_i^* \mid X))\ \mathrm{p}(T_i) / \textstyle\sum_i \exp(\ell_i(\theta_i^* \mid X))\ \mathrm{p}(T_i) \qquad \text{(I.19)}$$

where $\ell_i(\theta_i^* \mid X)$ denotes the log-likelihood function calculated at the MLE θ_i^*. When only two *randomly* chosen topologies are considered, Kishino and Hasegawa (1989) give a

means to estimate the 95% confidence interval of a given tree, approximating the log-likelihood function by a multivariate-normal distribution (asymptotic distribution of the sum of idd variables under the central limit theorem – e.g. Mood et al., 1974). This is the basis of what is known as the KH89 test.

This approach is similar to a more general approach developed in the statistical (Linhart, 1988) or econometrical (Vuong, 1989) literature. The general idea is the following. Given some regularity conditions, the asymptotic distribution of the LRT statistic can be obtained either when the two compared distributions are equal (case 1), or when they are not (case 2). Case 1 can be used to test whether there exist enough information in the data to distinguish different models (e.g Bar-Hen and Kishino, 2000); the deviance then follow a weighted sum of non-central χ^2 distributions, whose degrees of freedom can be calculated from the eigenvalues of the Fisher information matrix (Vuong, 1989). The relevance of case 2 appears when more specific hypotheses, such as whether one distribution is larger than another one, are tested. This can be used to select the candidate for the best tree (e.g Bar-Hen and Kishino, 2000). The standardised deviance now follows a normal distribution $N(0,1)$. When multiple comparisons are involved, as it happens when comparing more than two topologies, adjustments need to be done (e.g. Bonferroni, knowing it is not a good correction – see Chapter IV). The only application of these results to phylogenetics is given by Bar-Hen and Kishino (2000).

At the other extreme of the methods reviewed so far lie what can be denoted "non-parametric methods" to compare non-nested hypotheses. The bootstrap (Efron, 1979; Felsenstein, 1985; see also Efron and Tibshirani, 1993) is probably the most popular one, as it is easy to implement, and *apparently* intuitive (see below I.3.c). Let θ be a population parameter estimated from a sample $X = \{x_1,\ldots,x_n\}$ of n individuals: $\hat{\theta}$. The sampling distribution of θ is unknown, but we wish to estimate it. Assuming that all the

samples are equally probable, B random samples of size n are drawn with replacement from X. For each resampled data set X^*, the parameter $\hat{\theta}^*$ is estimated as in the original data set. The central idea is to use the observed distribution of the difference $\hat{\theta}^* - \hat{\theta}$ to infer the unobservable distribution of $\hat{\theta} - \theta$. In the phylogenetic problem, X is the vector of sites (columns of the data matrix where the x_i's are resampled with replacement until a data set X^* of the same size as X is obtained), from which a tree T is estimated: $\hat{T}$. This is repeated a large number B of times. The best (ML) tree is estimated for each replicate, and the frequency of appearance of each topology (or each tree bipartition) is recorded. This quantity is generally used to assess the relative stability of an estimated tree or of a bipartition, and is taken to represent the statistical significance at a given α level (e.g. Nei and Kumar, 2000, p. 172). This use of the bootstrap is however non-standard in that the statistic $\hat{T}$ is discrete, which creates part of the difficulties exposed below (see I.3.c).

Bootstrap probabilities, that is the "*frequency of a particular tree being the highest likelihood tree among alternatives during bootstrap resampling*" (Hasegawa and Kishino, 1994), are typically expensive to compute. This has motivated the search for approximations. Two methods have been proposed by Kishino et al. (1990). Noting that for a large number of site patterns the log-likelihood functions at their respective MLE approximately follow a multivariate normal distribution (MND), it is possible to estimate their variance-covariance matrix. This is the basis of the MND method, which aims at estimating the bootstrap probability that a selected tree is the best (Kishino et al., 1990). A more efficient method is to resample not the data matrix itself, but the estimated log-likelihood values for sites, hereby avoiding recalculating any likelihood function or its MND approximation. This method, known as RELL, which stands for "*r*esampling *e*stimated *l*og-*l*ikelihoods", as well as MND, have been shown to provide good approximations of bootstrap probabilities (Hasegawa and Kishino, 1994). I will come

back to the signification of these probabilities later on in terms of which null hypothesis is actually tested (Chapter IV).

More recently, some more tests have been developed and implemented (Swofford et al., 1996; Shimodaira and Hasegawa, 1999). See Goldman et al. (2000) and Whelan et al. (2001) for two reviews. A more detailed treatment will be given later (see Chapter IV).

Information-theoretic approaches

Akaike (1973) formulated what is probably the most well known of the model selection criteria, the Akaike information criterion (AIC). Let $Y = \{y_1, \ldots, y_n\}$ be observations from a sets of models, g_j, parameterised by θ_j. Each model g_j aims at approximating the true unknown distribution $f(\cdot)$. We have seen above that the K-L distance can be estimated by:

$$\hat{I}(f, g_j) = \int f(x) \log\{ f(x) / g_j(x \mid \theta_j^*(y)) \}\, dx \qquad \text{(I.20)}$$

where $\theta_j^*(y)$ is the MLE estimated from the data y under the model g_j. In the context of repeated sampling properties of an inference procedure, this estimated distance is expected to be $\int f(y)\{ \hat{I}(f, g_j)\} dy$. This can be written as:

$$\int f(y)\{ \hat{I}(f, g_j)\} dy = \text{constant} - E_y E_x[\log(g_j(x \mid \theta^*(y))] \qquad \text{(I.21)}$$

This later quantity, $E_y E_x[\log(g_j(x \mid \theta^*(y))]$, is estimated by the maximised log-likelihood $\log(g_j(x \mid \theta^*(y))$, which I now write $\ell_j(\theta_j^*)$. The reference to the true but unknown distribution $f(\cdot)$ drops out in the constant. The best approximating model, that is the one that has the shortest relative expected K-L distance to the true distribution, can in principle be selected as the one with the largest $\ell_j(\theta_j^*)$. However, Akaike (1973) showed that $\ell_j(\theta_j^*)$ is biased upward as an estimator of the model selection criterion based on the K-L distance. Under certain conditions, this bias is approximately p_j, the number of estimable parameters. *A* selection criterion is then approximately $\max_{g_j} [\ell_j(\theta_j^*) - p_j]$. For "historical reasons", Akaike (1973) defined AIC by:

$$\mathrm{AIC} = -2\ell_j(\theta_j^*) + 2p_j \qquad \text{(I.22)}$$

and the best approximating model is selected as $\min_{g_j} \mathrm{AIC}_j$. This approach has been used many times in phylogenetics (Hasegawa and Kishino, 1989; Kishino and Hasegawa, 1990a; Cao et al., 2000; Hasegawa et al., 1991; Posada and Crandall, 2001a). Note that, unlike what is stated by some authors (Nei and Kumar, 2000, p. 155), this approach is absolutely valid to select *non-nested* models, in particular tree topologies. However, it is known that AIC is not consistent (e.g. Woodroofe, 1982): in general AIC does not select the "correct" model as n tends to infinity. Several simulation studies (Dempster et al., 1977; Altman and Andersen, 1989) show that AIC tends to pick models which are too large (i.e. parameter rich). The recent results by Posada and Crandall (2001a) are consistent with these older simulation studies, but contrast with the results of a recent study which suggests that AIC tends to select simple models (Takahashi and Nei, 2000). This suggests that an alternative measure should be used.

Based on the assumption that a "true model" exists, and that it is one of the candidate models being considered, Schwarz (1978) introduced a consistent criterion. His Bayesian Information Criterion (BIC), defined as:

$$\mathrm{BIC} = \ell_j(\theta_j^*) - \tfrac{1}{2}\, p_j \log(n) \qquad \text{(I.23)}$$

aims at selecting this "true model". Note however that BIC is not directly linked to the K-L information and is "information-theoretic" only in the weakest sense. Often viewed as a penalised version of AIC, BIC leads to a correct choice of model as n tends to infinity (Haughton, 1988). This idea of penalising AIC produced other criteria, one of which has recently been used in phylogenetics to explore the demographic history of sampled DNA sequences (Strimmer and Pybus, 2001). Standard Bayesian testing procedures use the Bayes factor (Kass and Raftery, 1995, see II.1.d and IV), which is the ratio of posterior to prior odds under two competing models. Kass and Raftery (1995) showed that BIC can

be derived via a Laplace approximation to the Taylor expansion of the Bayes factor (BF) between two models. BIC approximates the logarithm of the Bayes factor, and satisfies:

$$\lim_{n\to\infty} \{(\log(BF) - BIC) / \log(BF)\} = 0 \qquad (I.24)$$

However:

$$\lim_{n\to\infty} \{\exp(BIC) / BF\} \neq 1 \qquad (I.25)$$

More precisely, log(BF) – BIC has asymptotic error O(1). Therefore, although BIC can be viewed as a rough approximation to the logarithm of BF, the error associated with this approximation does not vanish as n gets large so that BIC is not a fully consistent approximation of log(BF), at least for some prior distributions. Kass and Wasserman (1995) showed that under a different choice of prior, based on Fisher information matrix, BF can be asymptotically approximated by exp(BIC), the ratio exp(BIC) / BF tending to one with an error $O(n^{-1/2})$. This implies that, although BIC is a procedure which does not require the specification of a prior, it approximates a Bayes factor which is based on a particular prior for the parameter of interest. Therefore, when using BIC to assess models, an implicit prior, called the overall unit information prior, is used. Moreover, the value of n in the penalty term of BIC is not always obvious. This is similar to calculating the degrees of freedom when comparing topologies in a ML framework (Goldman, 1993; Yang et al., 1995).

For these reasons, it appears more reasonable to avoid any approximation and select the best approximating model by means of the Bayes factor. In many cases, calculating a Bayes factor entails calculating complex integrals which generally have no closed form solutions. Markov chain Monte Carlo integration (Chapters II and IV) offers a solution to approximate these integrals. This solution is generally not the one favoured (Kass and Raftery, 1995), but appeared to be stable enough in our case for practical and efficient model selection.

I.3.c – More about the non-parametric bootstrap

As seen above (I.3.b), the bootstrap *p*-values (p_B) are usually taken as a measure of support for the presence of groups in a phylogeny, i.e. that this group is monophyletic. Almost a decade ago, Zharkikh and Li suggested, on the basis of a theoretical argument (Zharkikh and Li, 1992a) and of simulations (Zharkikh and Li, 1992b), that p_B is a very conservative measure to estimate the reliability of an inferred phylogeny. In particular, it tends to underestimate the true *p*-value when this latter is high, and overestimate it otherwise. This result was established in the four-taxa case, and extended to a larger number of taxa by another simulation study (Hillis and Bull, 1993), where the authors suggested disregarding p_B altogether. From this point on, new interpretations of p_B have been proposed. For Felsenstein and Kishino (1993), $1 - p_B$ is the probability of type I error, that is the probability of rejecting the null hypothesis that the tree is multifurcating when the tree is actually bifurcating. Under this interpretation, $1 - p_B$ is correct only when the null hypothesis is true; otherwise, it is conservative (Felsenstein and Kishino, 1993).

This interpretation led to the interior branch test for trees estimated by distance methods (Sitnikova et al., 1995). In this test, the expected branch lengths are positive or null under the true topology, but the other trees have at least one such expectation that is negative. Dopazo (1994) proposed a bootstrap version of this analytical procedure to test the non-negativity of interior branch lengths. However, this is difficult to apply to ML, where all the branch lengths are positive. Felsenstein (1988) suggested that the null hypothesis that one of the interior branch length b_i is zero be tested in a ML framework. However, while $b_i = 0$ is nested in $b_i \geq 0$, the null hypothesis specifies a parameter at the boundary of the parameter space, so that the deviance follows not a χ^2 distribution (Gaut and Lewis, 1995), but a simple mixture (Self and Liang, 1987; see I.3.a). Although Gaut

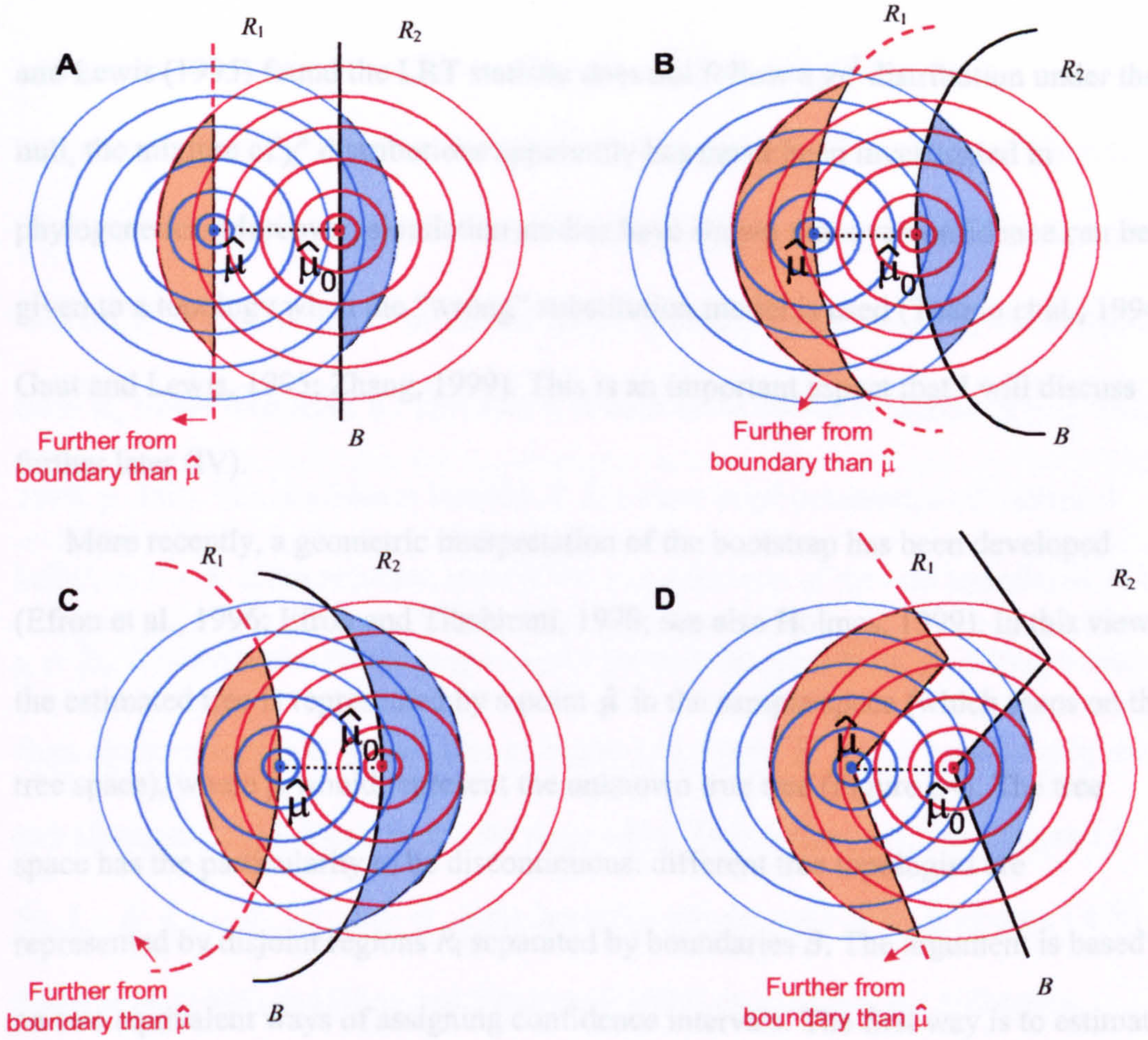


Figure I.4. What is the shape of the boundary between two topologies in the tree space? **A** linear? **B** convex (the boundary B curves away from $\hat{\boldsymbol{\mu}}$)? **C** concave (the boundary B curves towards $\hat{\boldsymbol{\mu}}$)? **D** non-smooth? The bootstrap density of $\hat{\boldsymbol{\mu}}^*$ is represented by areas comprised within circles around the estimated vector $\hat{\boldsymbol{\mu}}$. This point lies in the region R_1 of the estimated ML tree. The closest other tree is in the region R_2, separated from R_1 by the boundary B. See text for details.

and Lewis (1995) found the LRT statistic does not follow a χ_1^2 distribution under the null, the mixture of χ^2 distributions apparently has never been investigated in phylogenetics. Moreover, simulation studies have shown that overconfidence can be given to a topology when the "wrong" substitution model is used (Tateno et al., 1994; Gaut and Lewis, 1995; Zhang, 1999). This is an important aspect that I will discuss further later (IV).

More recently, a geometric interpretation of the bootstrap has been developed (Efron et al., 1996; Efron and Tibshirani, 1998; see also Holmes, 1999). In this view, the estimated tree is represented by a point $\hat{\mu}$ in the sample space (which maps on the tree space), where μ would represent the unknown true tree (Figure I.4). The tree space has the particularity to be discontinuous: different tree topologies are represented by disjoint regions R_i separated by boundaries B. The argument is based on two equivalent ways of assigning confidence intervals. The first way is to estimate the distribution of $\hat{\mu} - \mu$ by bootstrapping the original data. The distribution of $\hat{\mu}^* - \hat{\mu}$ is obtained, where $\hat{\mu}^*$ represents the trees estimated from the bootstrapped data (Felsenstein, 1985; Efron et al., 1996). The confidence level $\tilde{\alpha}$ for the event that the true tree μ lies in region R_1 is measured by the distance of $\hat{\mu}$ from the boundary (Figure I.4 – resampling, i.e. the bootstrap density of $\hat{\mu}^*$, is indicated by the blue concentric circles). Stated differently, the rejection level $1 - \tilde{\alpha}$ is measured by the proportion of bootstrapped replicates from $\hat{\mu}$ falling in R_2 (blue area in Figure I.4). This approach has a Bayesian interpretation, as $\tilde{\alpha}$ is the posterior probability that μ lies in region R_1 given that $\hat{\mu}$ lies in region R_1. By assuming a uniform prior on the different topologies in the tree space, we have:

$$\tilde{\alpha} = \int_{x^* \in R_1} p(\hat{\mu}^* \mid \hat{\mu})\, dx^* \qquad \text{(I.26)}$$

which is the bootstrap proportion traditionally computed in phylogenetics (Felsenstein, 1985). Another customary way of assigning confidence intervals is to use the point $\hat{\boldsymbol{\mu}}_0$ that is closest to $\hat{\boldsymbol{\mu}}$ and located on the boundary B. The confidence interval $\tilde{\alpha}$ is then measured by the distance from the boundary to $\hat{\boldsymbol{\mu}}$, by bootstrapping from $\hat{\boldsymbol{\mu}}_0$ rather than from $\hat{\boldsymbol{\mu}}$. This idea is related to the notion of similarity (e.g. Shao, 1999, p. 356). The confidence interval $\hat{\alpha}$ is a more usual assessment of statistical belief, as $1 - \hat{\alpha}$ is the rejection level of the one-sided test of the null hypothesis that $\boldsymbol{\mu} \notin R_1$. It is measured by the proportion of points resampled from $\hat{\boldsymbol{\mu}}_0$ falling further from the boundary than $\hat{\boldsymbol{\mu}}$ (red area in Figure I.4). While $\tilde{\alpha}$ and $\hat{\alpha}$ measured by the two aforementioned distances are the same when the boundary is linear (Figure I.4 **A**), $1 - \tilde{\alpha} < 1 - \hat{\alpha}$, i.e. $\tilde{\alpha} > \hat{\alpha}$, if the boundary curves away from $\hat{\boldsymbol{\mu}}$ (Figure I.4 **B**). Efron et al. (1996) and Efron and Tibshirani (1998) showed that the difference between $\tilde{\alpha}$ and $\hat{\alpha}$ amounts to choosing a different prior distribution for $\tilde{\alpha}$. Assuming the boundary is a smooth curve, these authors showed that setting a uniform prior in polar coordinates on the distribution of topologies in the tree space reconciles both confidence measures. However, if the curvature of the boundary is rugged (Figure I.4 **D**), as suggested by Yang (2000a, see his Figure 4), further complications are likely to arise: fitting a smooth curve to measure the distance further from the boundary than $\hat{\boldsymbol{\mu}}$ in Figure I.4 **D** generates a bias (under- or overestimation of $\hat{\alpha}$, depending on the exact geometry of the boundary).

It is clear from this ongoing debate (i) that the interpretation of the bootstrap support for internal nodes is still not simple and (ii) that tests developed to date may not estimate correctly the confidence in an estimated tree. Conservative approaches, advocated by Nei and Kumar (2000, p.174), may not always be appropriate.

To conclude this first chapter, I would like to emphasise two critical points. The first relates to estimating parameters that are confounded. The typical example is that of estimating times *and* rates of evolution, without assuming which sets of branches have equal rates (molecular clocks, global or local). The second issue is that of testing hypotheses, and in particular testing phylogenetic trees. We have seen, and will delve with more details in Chapter IV, that it is possible to build confidence sets of "equally good" trees or to answer questions such as "which tree is the correct tree?". This very latter point typically demands awkward spectral decomposition to compute complicated degrees of freedom involved in equally complicated mixtures of χ^2 distributions. Moreover, the issue of multiple comparisons is most of the time not dealt with properly. The Bayes approach I present in the next chapters is an attempt to give intuitive answers to complicated questions. Chapters II and IV will mainly focus on the theory and the validity of the argument in face of ML methods, while Chapters III and IV shall provide the reader with some examples of real data analysis.

Chapter II

–

Bayes inference of times and rates: the model and its implementation

II.1 – The Bayesian approach

Since it was proposed by Zuckerkandl and Pauling (1965) almost four decades ago, the molecular clock hypothesis, that is, the constancy of evolutionary rate over time, has been a matter of debate (Chapter I). A number of tests have been developed to examine its validity, such as the relative rate test (Sarich and Wilson, 1967; Wu and Li, 1985) and the likelihood ratio test (Felsenstein, 1988). These tests often reject the molecular clock in real data sets (see Nei and Kumar, 2000, p.188). Since the rate of evolution is not constant across lineages, it is interesting to know whether it fluctuates at random, or evolves following some specific trends.

Recently, several studies have attempted to relax the molecular clock assumption when estimating divergence times. One approach is to construct local molecular clock models in the likelihood framework (see section I.2.c), where independent evolutionary rates are assigned to some lineages while all the other branches evolve at the same rate. Dates and rates are then parameters in the model and are estimated by maximum likelihood (ML). This approach is straightforward to apply if the branches with different rates can easily be identified a priori. However, when such information is unavailable, date estimates might be sensitive to the assumptions about the rates.

Another approach is to use a stochastic process to describe evolutionary rate change over lineages, relying on the observation that closely related lineages tend to have similar rates (Sanderson, 1997, 2002). A Bayesian approach is then used to derive the posterior distributions of rates and dates. One such model assumes that rates are autocorrelated across speciation events: the rate of a branch is sampled from a lognormal distribution centred on the rate of the ancestral branch (Thorne et al., 1998). Other models of rate evolution have also been suggested. Following Gillespie (1991), Bickel (2000) and Cutler (2000a) proposed a model based on a doubly stochastic Poisson process (Cox process),

which extends the constant-rate Poisson process first described to model the accumulation of substitutions since divergence (Zuckerkandl and Pauling, 1965). Huelsenbeck et al. (2000) modelled the evolutionary rate as a point process, assuming that the rate of evolution changes according to a Poisson process along the tree, while the rate parameter of the Poisson has a gamma prior distribution.

There seem to be some arbitrariness in the choice of the model of rate evolution. Thus, it is important to know how sensitive the estimates of divergence times are to the choice of the model of rate change. In this chapter, I implement and compare different models of autocorrelated rate change over time and focus on two points: the effect of the model of rate change and the effect of the parameterisation of each model to relax the clock. I also compare these Bayesian methods with the likelihood-based local clock analysis. I use the hominoid tRNA gene (Horai et al., 1992) and the metazoan 18S rRNA gene (Bromham et al., 1998) as test data sets.

II.1.a – Bayesian modelling of evolution of times and rates

In the framework of maximum likelihood, the most general model assumes that the substitution rate r_i for branch i is allowed to vary among branches. The branch length is given by the product of the rate and the time duration for that branch, $b_i = r_i t_i$. The likelihood, that is, the probability of observing the data X, depends on the vector of branch lengths B and is denoted $p(X|B)$. Branch lengths can be estimated using classical hill-climbing algorithms to maximise the likelihood (e.g. see Gill et al., 1981). As rate and time are confounded, one cannot estimate one without making assumption(s) regarding the other. For instance, the molecular clock hypothesis assumes that all rates are equal; branch lengths are then proportional to divergence times, and the problem reduces to ML estimation. Models of local clocks are similar: some pre-specified branches are assigned independent rate parameters while all other branches have the same rate.

To relax the molecular clock in a Bayesian framework, a prior distribution $p(R, T)$ for rates of evolution R and divergence times T is chosen. The Bayes theorem is then used to derive the (posterior) probability of times and rates:

$$p(R,T \mid X) = \frac{p(X \mid R,T)p(R,T)}{p(X)} \qquad \text{(II.1)}$$

A sensible way to factorise the joint prior distribution is $p(R, T) = p(R \mid T)\, p(T)$. It is therefore assumed that speciation events are generated by a random process independent of the rates of molecular evolution and that the rate for a given branch is dependent on the time duration of that branch. Moreover, if the prior for the rates is independent of divergence times, it gives $p(R \mid T)\, p(T) = p(R)\, p(T)$.

The probability $p(X \mid R, T)$ is the traditional likelihood, and its calculation requires a nucleotide substitution model. In this chapter, the HKY85 model (Hasegawa et al., 1985) incorporating among-site rate variation modelled by a gamma distribution (Yang, 1994b) is used. The parameters in the substitution model are $\psi = \{\kappa, \alpha, \pi\}$, where κ is the transition to transversion rate ratio, α is the shape parameter of the gamma distribution and π is the vector of the base frequencies. This model is extended to take into account heterogeneous site partitions in the sequence (e.g. the three codon positions of a gene). Usually ψ is assumed to follow a uniform prior distribution with its components mutually independent and independent of R and T. The prior distribution of the complete model is then $p(R \mid T)\, p(T)\, p(\kappa)\, p(\alpha)\, p(\pi)$. In this section, I will concentrate on prior models for times and rates, with ψ set to its ML estimates (MLEs) obtained without the clock.

II.1.b – Prior distributions for divergence times

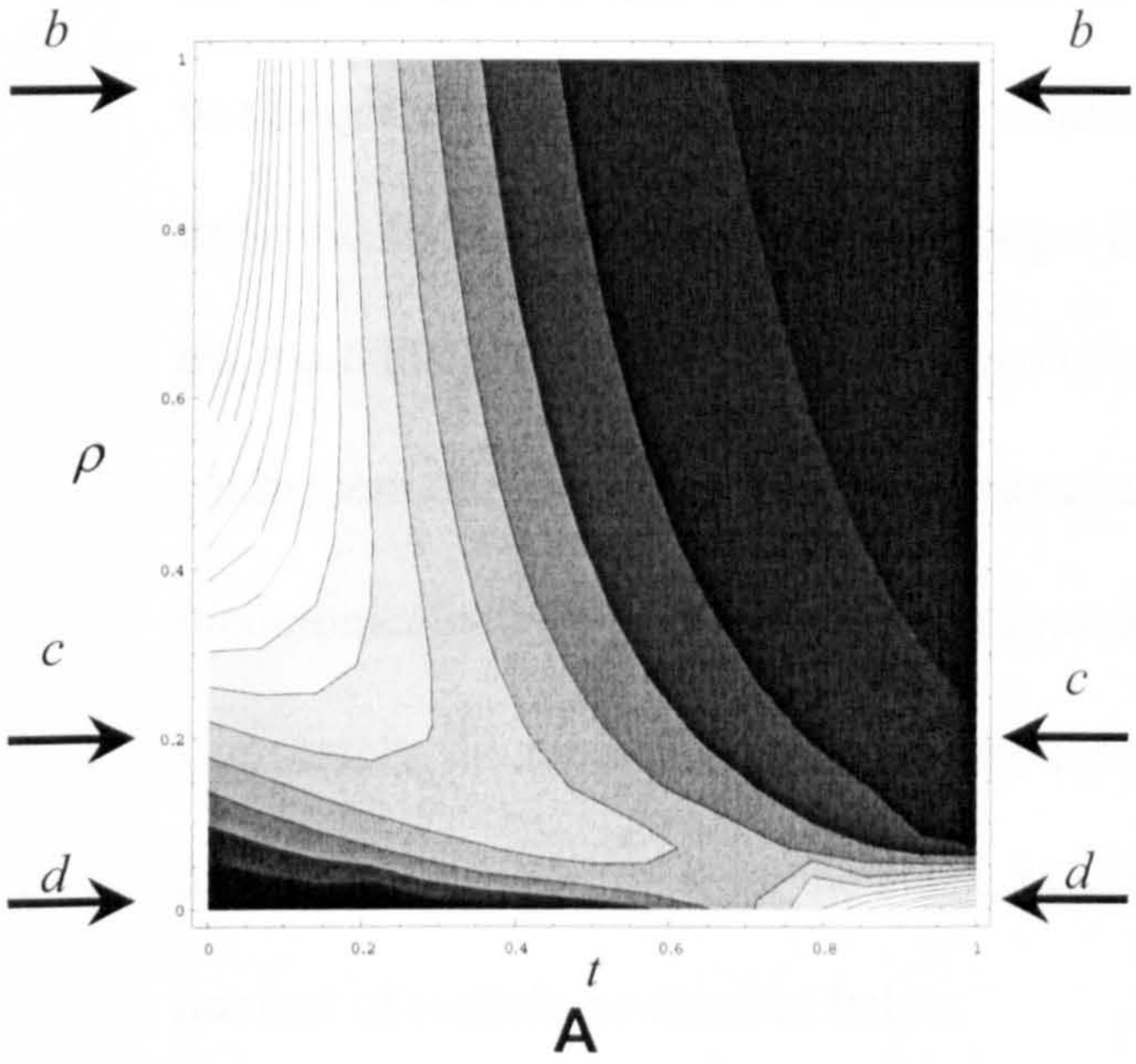
The prior distribution for divergence times is generated by a process of cladogenesis, the generalized birth and death process (BDP) with species sampling, as described by Yang

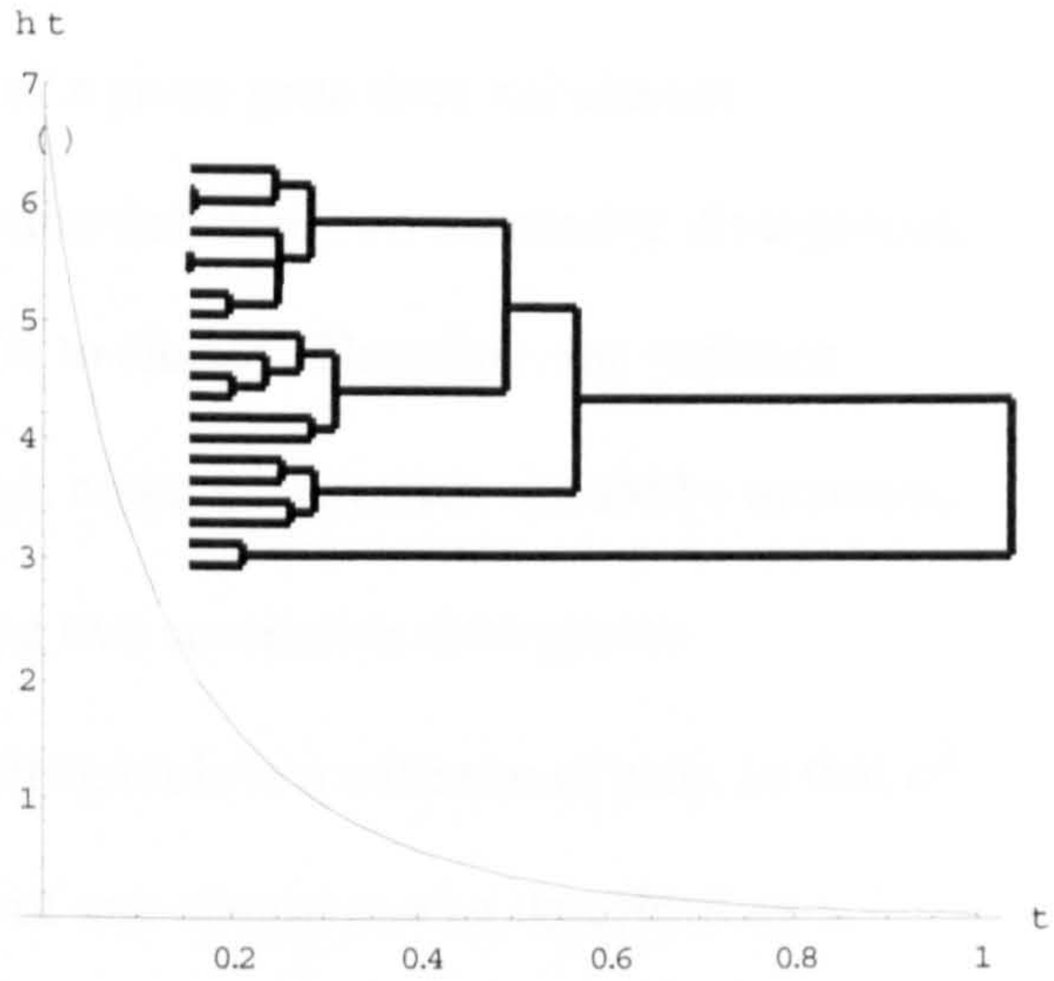
and Rannala (1997); see also Kendall (1948), Thompson (1975) and Nee et al. (1994). The model assumes a constant speciation rate λ and extinction rate μ per lineage. Node times are conditioned on the time of the root, arbitrarily set to one. Species sampling is modelled as a mass-extinction event occurring at the sampling time with a probability ρ. This process is flexible and can accommodate more shapes of trees than the Yule process as used by Thorne et al. (1998). In particular, taking the incomplete species sampling into account allows the distribution of divergence times to take any shape between L-shaped and J-shaped distributions (Figure II.1). However, this process seems comparable with the generalized Dirichlet distribution of Kishino et al. (2001). In order to accommodate the uncertainty in the hyperparameters (λ, μ, and ρ), they are integrated out of the model by a standard Bayes averaging method. Independent uniform distributions were used as priors for λ, μ, and ρ.

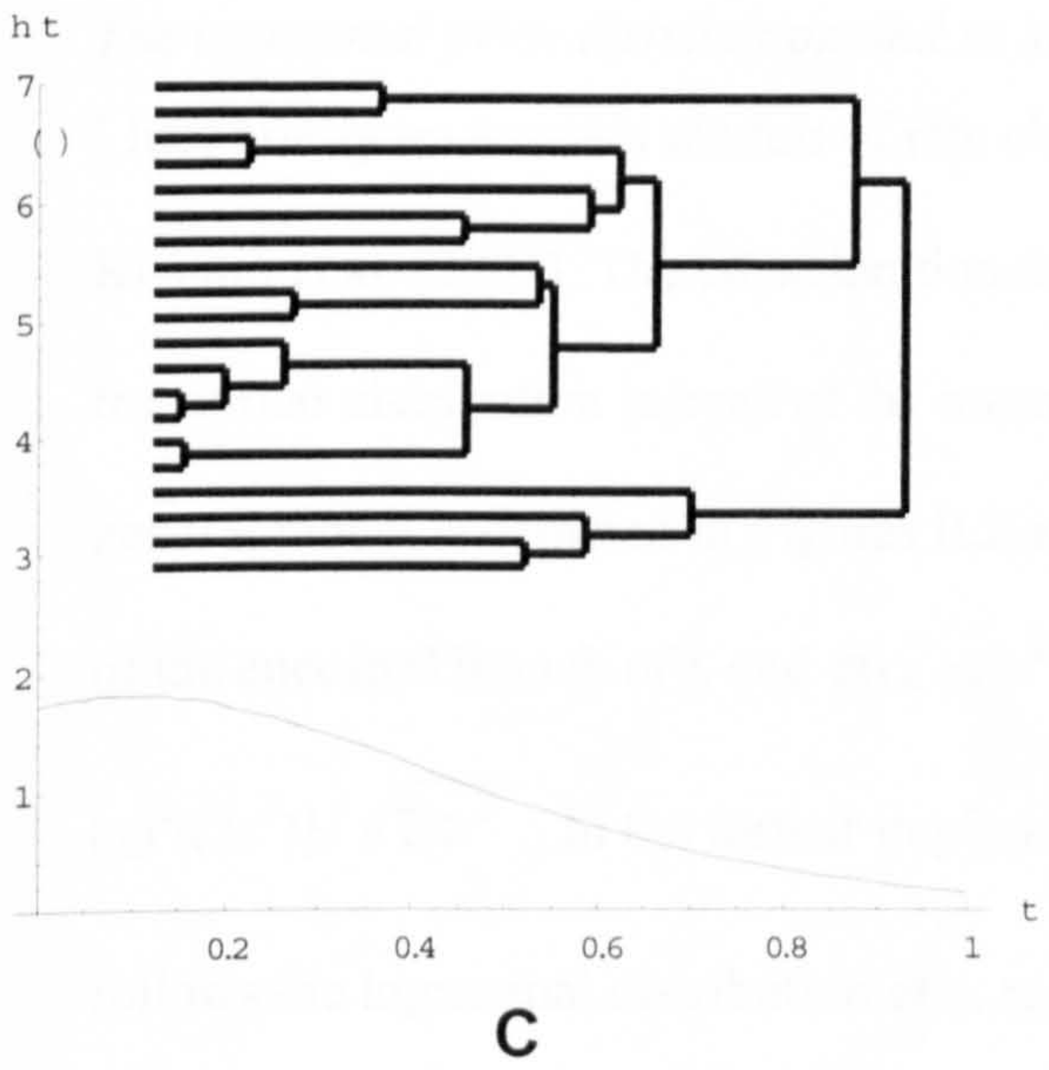
It is possible to incorporate lower and upper bounds on node times from fossil dates. This is expected to improve convergence of the algorithm, since the times for the constrained nodes do not have to explore the whole sample space. This has been implemented recently by Kishino et al. (2001), but as pointed out by those authors, no appropriate prior under such constraint has been suggested. As a result, this feature is not incorporated in the current implementation.

II.1.c – Prior distributions for rates of evolution

The models of rate change presented and developed here are based on the following two ideas. First, we assume that the evolution of the rate of molecular evolution over the time separating two nodes of a tree can be described by considering a branch-specific rate, which represents the mean rate over this time period. The change of such a branch-specific rate is then described by a statistical process, whose mean is centred on the rate of the direct ancestor. Hence, we do not assume any trend in rate evolution, either upward or as a







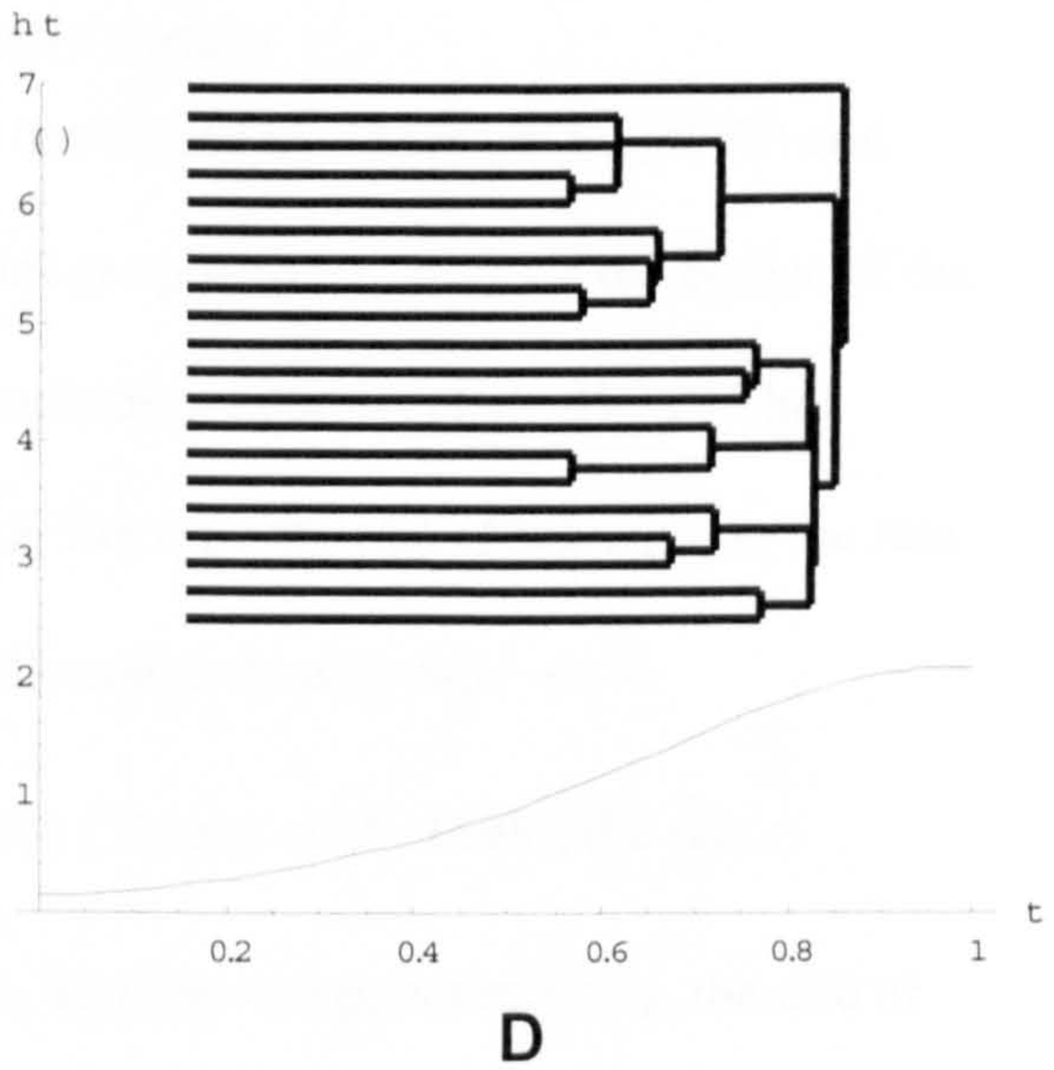


Figure II.1. Effect of incomplete species sampling on the distribution of divergence times under the generalised birth-death process. **A** Contour plot of the pdf of the divergence times (t) as a function of the sampling fraction (ρ) assuming fixed birth rate ($\lambda = 6.7$) and death rate ($\mu = 2.5$); times regions of lower density are of darker shades. Arrows indicate cross sections represented in the other panels of this Figure. **B** Cross-section of panel (A) corresponding to $\rho = 1.0$. **C** Cross-section of panel (A) corresponding to $\rho = .25$. **D** Cross-section of panel (A) corresponding to $\rho = .01$.

slowdown. Second, if two divergences occur during a relatively short time period, it can be thought that the rate of molecular evolution of a given gene does not change dramatically. On the other hand, the longer the time between two successive divergences, the more likely the rate of molecular evolution is to change. Therefore, the variance component of the process describing rate change, noted σ^2 hereafter, should be increased monotonically with the time period Δt separating two successive divergences. Conversely, the molecular clock assumption corresponds to a variance of zero, so that σ^2 measures the departure from the clock. Models of rate change can be described by a number of models, as detailed below.

The lognormal prior distribution and its stationary variant

I first briefly review two models of rate change developed by Thorne et al. (1998) and Kishino et al. (2001). The consideration that rates are positive motivated the choice of the lognormal distribution instead of the normal distribution (Thorne et al., 1998). The general model is depicted in Figures II.2 and II.3. Let r_i be the rate of branch i, r_A the rate of the ancestral branch of i, and $\varphi(r_i, r_A, s^2)$ the Gaussian density $\exp\{-(r_i - r_A)^2/(2s^2)\}/\sqrt{2\pi s^2}$. In the former implementation (Thorne et al., 1998), the rate r_i follows the lognormal distribution $\varphi(r_i, r_A, s^2)/r_i$, and has two parameters: r_A, the rate of the ancestor, and s^2, a variance parameter that controls how much the model is constrained by the clock. This lognormal model is hereafter referred to as LND. If the time period between two speciation events is short, it is natural to think that the rate of evolution of a given gene may not change dramatically. On the other hand, the longer this period, the more likely the rate changes. Therefore, s^2 was assumed to be proportional to this time period, Δt, with $s^2 = \sigma^2 \Delta t$. Parameter σ^2 measures the departure from the strict clock assumption: the model tends to the molecular clock for small σ^2, and represents highly variable rates when σ^2 is large. The time duration Δt was measured by the

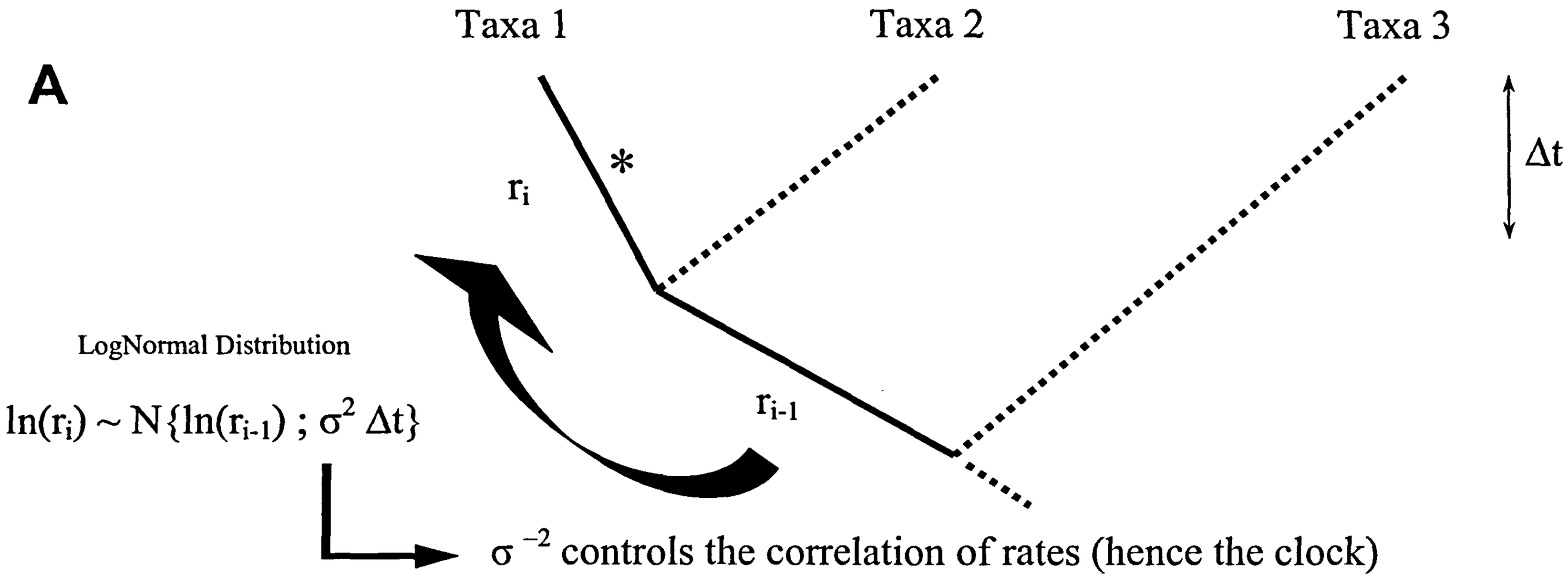


Figure II.2. Schematic representation of the general model of autocorrelated rate change. **A** The model for any branch of the tree except the root (see text for details). **B** Case of the root: as the root has no ancestral rate, updating one of the branches emanating from the root demands a special treatment. The time duration Δt is taken as the one of the current state in the Markov chain, while the "ancestral" rate is taken as the average of the rate of these two branches at the current state.

difference between the two midpoints of the current and ancestral branches in Thorne et al. (1998) and by the time duration of the current branch in Kishino et al. (2001).

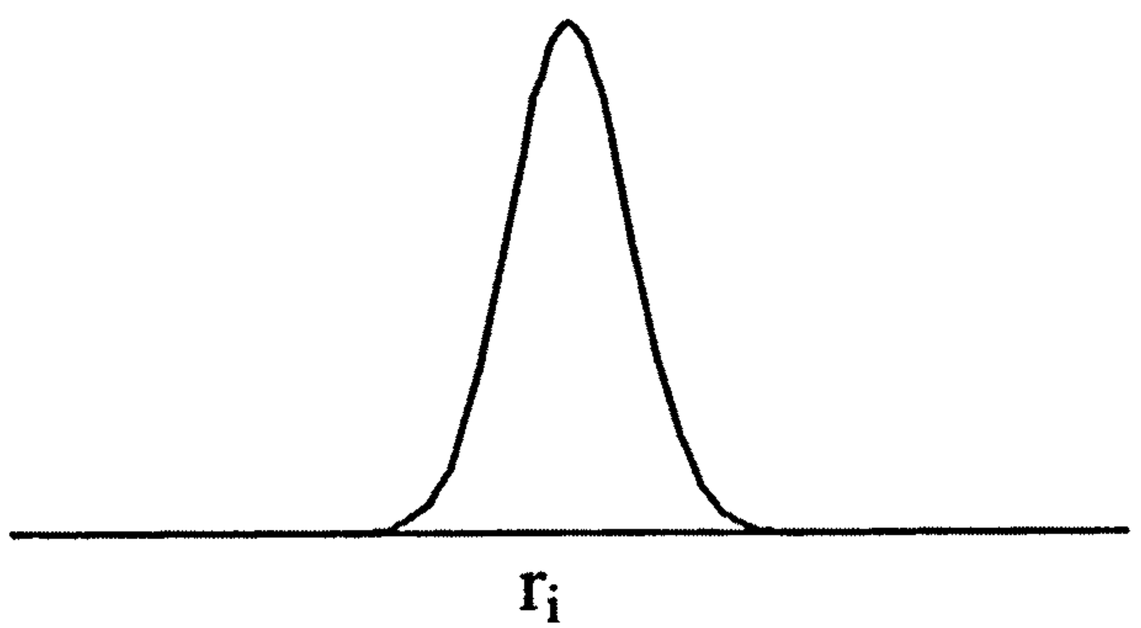
Note that the mean of the lognormal distribution is not the ancestral rate r_A, but $r_A\, e^{s^2/2}$. The rate of evolution therefore exhibits an upward trend, and so the process is time-dependent. A remedy to this problem, proposed by Kishino et al. (2001), is to subtract $s^2/2$ from the logarithm of the ancestral rate, so that the probability density function (pdf) becomes $\varphi(r_i, r_A\, e^{s^2/2}, s^2)/r_i$. I refer to this modified distribution as the stationary lognormal distribution (SLD).

To reduce the computational demand, Thorne et al. (1998) and Kishino et al. (2001) used MLEs of branch lengths $\hat{B}$ as pseudo-data, approximating the likelihood function by a multivariate Normal distribution centred on $\hat{B}$. My implementation of the LND and SLD models is similar to those of the previous authors (Thorne et al., 1998; Kishino et al., 2001), but I adopted an exact and more expensive likelihood computation using the sequence alignment.

The gamma and the exponential distributions

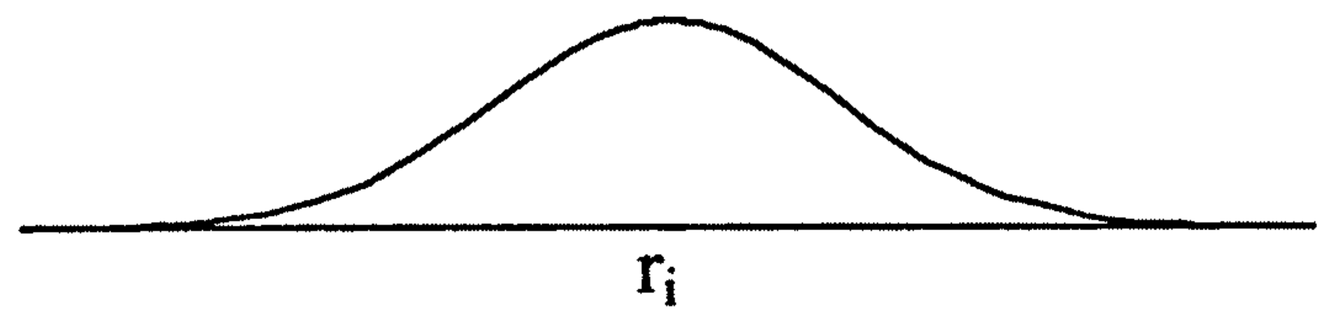
I also implemented two simple models of rate change: the gamma and the exponential distributions, referred to as GD and ED respectively. The rate of a branch is assumed to be drawn from a gamma or an exponential distribution, with the mean rate equal to the rate of the ancestral branch. As with the models discussed above, the variance of the GD is set proportional to Δt, the time duration of the considered branch, whereas with the ED, the variance is a function of the mean only. Therefore, these two models are not nested, and ED implicitly assumes that the larger the rate, the more variable it is.

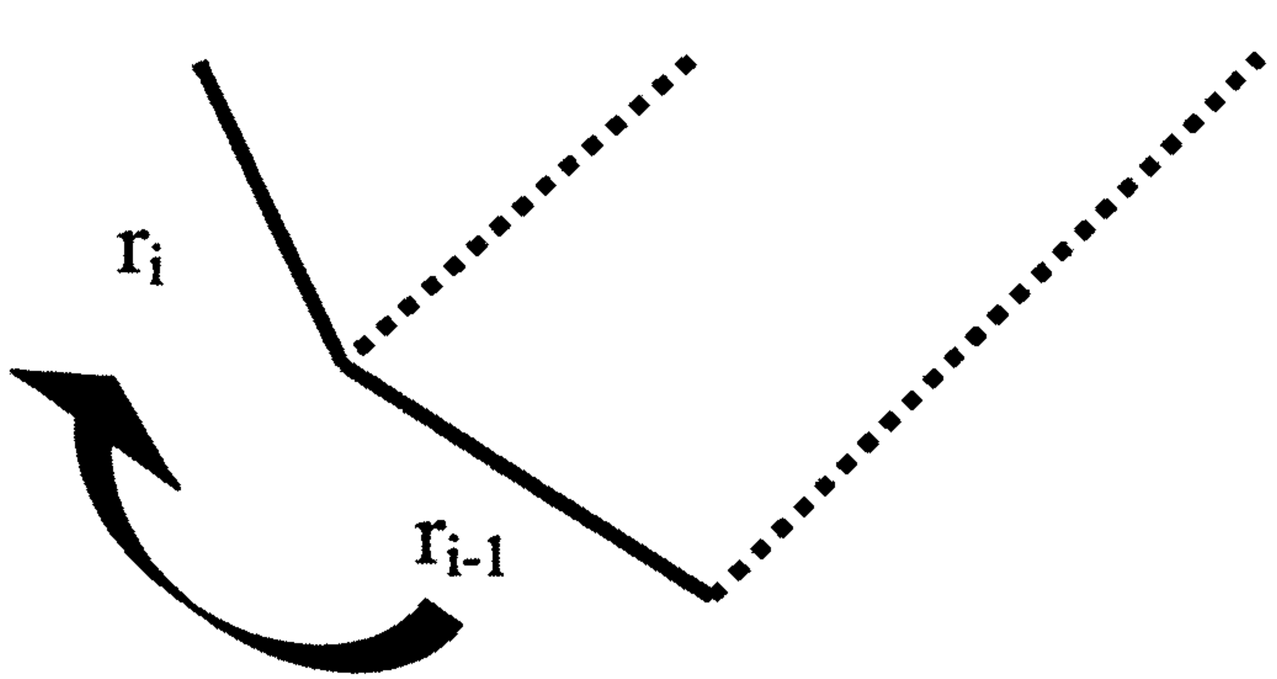
correlated rates

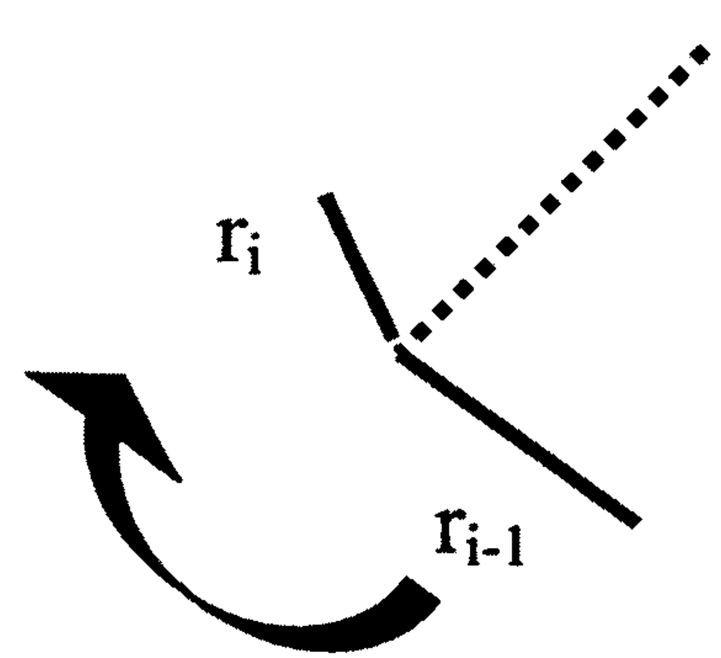


B large σ^2: loosely correlated rates

distribution of rates
(r_{i-1})







Clock-like tree

Figure II.3. The role of the variance in the autocorrelated process modelling rat

The Ornstein-Uhlenbeck process

Another implemented model allows the rate to evolve according to the Ornstein-Uhlenbeck process (noted OUP), a time-continuous Gaussian Markov process. OUP was originally designed to model the speed of a particle (and not just its position as in the Brownian process) as a function of time. The speed of the particle is reduced by frictional resistance from the medium and altered by random collisions with neighbouring particles. According to the process, the pdf of rate r_i is $\varphi(r_i, r_A e^{-\beta\Delta t}, \sigma^2(1 - e^{-2\beta\Delta t})/(2\beta))$ (e.g. Karlin and Taylor, 1981, pp. 170-173; Cox and Miller, 1965, p. 226). The mean of the distribution is now $r_A e^{-\beta\Delta t}$, which tends to the ancestral rate r_A as the hyperparameter β and/or Δt go to zero. As before, σ^2 is the parameter measuring departure from the molecular clock: the variance of the distribution, $\sigma^2(1 - e^{-2\beta\Delta t})/(2\beta)$, tends to $\sigma^2\Delta t$ for small β and/or small Δt.

Finally, note that the simplest model of rate change is when all the branches of the tree have the same rate. This is essentially the Bayesian version of the molecular clock hypothesis, the only difference with the traditional clock being the prior distribution for the speciation times.

II.1.d – Prior model selection

Our primary interest is to estimate divergence dates. However, different models of rate change can lead to different date estimates, and choosing the model that best fits the data can be critical. In a Bayesian framework, inference proceeds usually from the posterior distribution $p(\theta|X)$, where θ stands for parameters R, T and the hyperparameters of the birth-death process. However, $p(\theta|X)$ does not allow us to evaluate the goodness of fit of the model, nor does it permit comparison between models, which have different sets of parameters.

The marginal probability $p(X)$ under a given model M_k, also denoted $p(X|M_k)$, contains information for assessing model performance. One approach is to use the Bayes factor to compare models M_1 and M_2: as $BF_{12} = p(X|M_1) / p(X|M_2)$. Note that $p(X)$ for each model is obtained by averaging (and not maximizing, as for the likelihood ratio test) over the parameter space, with respect to the prior distribution. The so-called prior mean is defined as:

$$p(X \mid M_k) = \int p(X \mid R,T)\ p(R,T \mid \eta)\ p(\eta)\ dR\ dT\ d\eta \qquad \text{(II.2)}$$

where η includes the hyperparameters from the birth-death process $\eta = (\lambda, \mu, \rho)$. Computing the right hand side of equation (II.2) is difficult (Raftery, 1996). Instead of computing the prior mean, Aitkin (1991) proposed to use the posterior mean under each model, which is used here as it can be calculated easily by sampling from the MCMC, i.e. with respect to the posterior distribution $p(R, T, \eta \mid X)$:

$$L_k^{post} = \int \{p(X \mid R,T)\ p(R,T \mid \eta)\}\ p(R,T,\eta \mid X)\ dR\ dT$$

$$L_k^{post} = E_{post}[p(X \mid R,T)\ p(R,T \mid \eta)] \qquad \text{(II.3)}$$

so that the posterior Bayes factor PBF_{12} comparing models M_1 and M_2 is L_1^{post} / L_2^{post}.

The posterior Bayes factor was criticised for using the same data twice (see discussion to Aitkin, 1991): first to obtain the posterior distribution of $\theta = (R, T, \eta)$ and second, to average the likelihood with respect to the posterior (using θ sampled from the posterior). Such a "double use of the data" entails a lack of coherence. In particular, the method was originally proposed to reduce the weight of prior assumptions in model comparison. However, using the data twice amounts to considering a highly informative prior, concentrated around the posterior distribution. This potentially flaws model comparison results, as shown by Cox and by Goldstein (in discussion to Aitkin, 1991). A remedy would be to partition the data as $X = (X_1, X_2)$, using X_1 to compute the posterior

distribution and X_2 to select the model. However, it is not sure whether this can be reliable for sequence data sets with non-iid columns (e.g. because of among-site rate variation).

Here, I have used the posterior Bayes factor principally because it appeared computationally more stable than the harmonic mean estimator of $p(X)$ to estimate the prior mean (not shown; Raftery, 1996; but see Chapter IV). While *PBF* may not be coherent, in the applications given below it favoured models of rate change over the molecular clock assumption, which is consistent with the LRT of the clock (see II.2 and Chapter III). Similarly, OUP was favoured as a model of rate change over the unrealistic EXP model in 19 out of 22 genes sampled (Chapter III). More importantly, the results presented below are insensitive to the model of rate change used, so that conclusions are not critically dependent on the *PBF*.

Note that in the present Chapter the uncertainty over η is not integrated out: the object here is to evaluate some of these hyperparameters. For the LND, SLD, GD, and OUP models of rate change, an empirical Bayes approach is used to estimate the hyperparameter σ^2. Under each model, L_k^{post} is evaluated for different values of σ^2. The value with largest *PBF* is chosen as the estimate. The same approach is used for the hyperparameters σ^2 and β under OUP.

II.1.e – The posterior distribution and its approximation

In a Bayesian framework, the marginal posterior distribution of a variable is obtained by integrating out other variables. For example, the marginal posterior distribution of the times T is derived from equation (II.1) by integrating $p(R, T|X)$ over the rates and the hyperparameters:

$$p(T|X) = \int \frac{p(X|B)p(R|T,\sigma^2)p(T|\lambda,\mu,\rho)p(\lambda,\mu,\rho)}{p(X)} dR\, d\lambda\, d\mu\, d\rho \qquad \text{(II.4)}$$

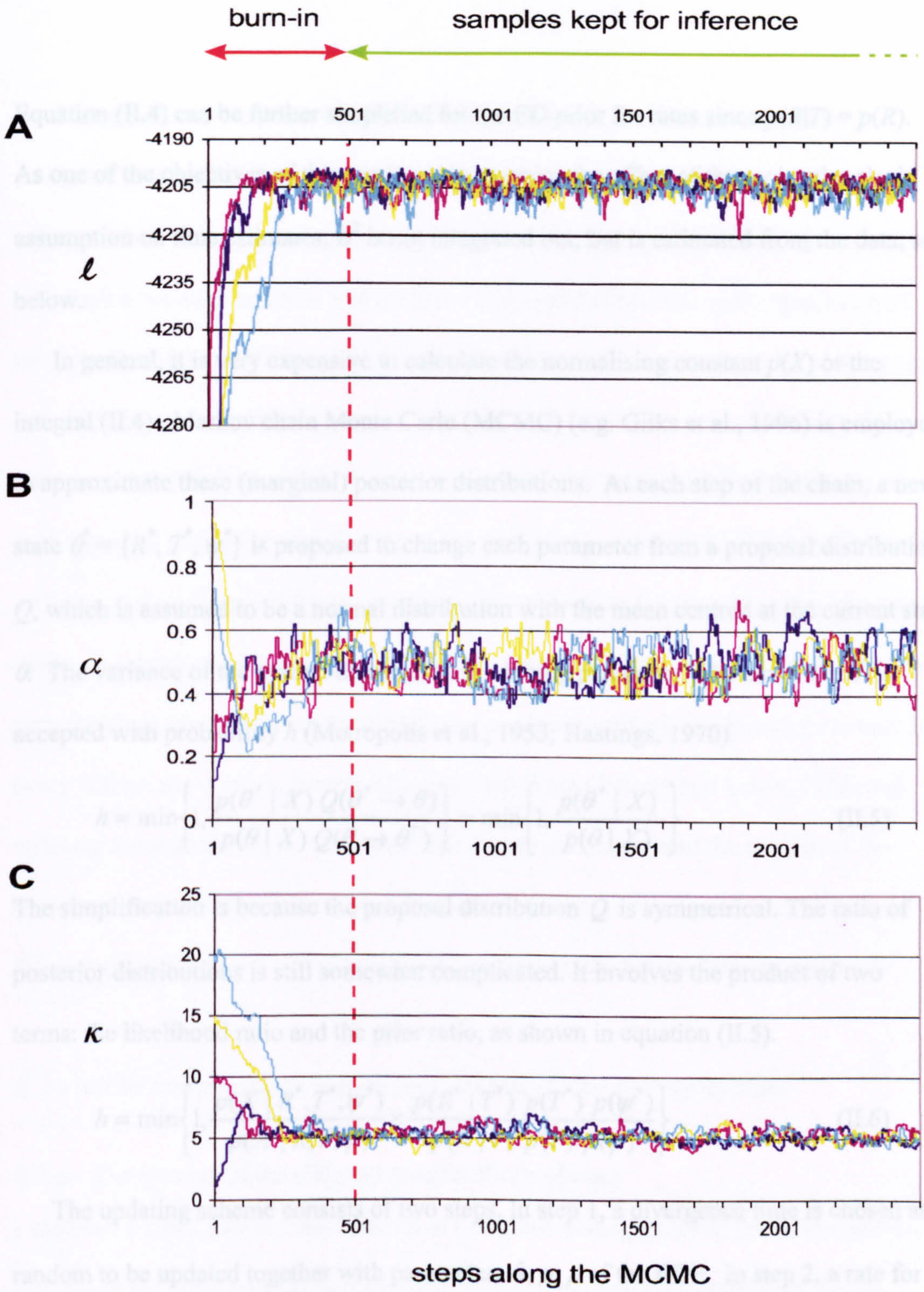


Figure II.4 Example of monitoring convergence and sampling along the MCMC for a simulated data set. The sequences 1,000 bp long were simulated for six taxa for the tree ((((1:0.1, 2:0.1):0.1, 3:0.2):0.1, 4:0.2), (5:0.1, 6:0.1):0.1). The HKY85 + Γ was used (κ_{sim}=5; α_{sim}=.5; $\pi_i=\pi_j$, $\forall\, i$ and j). MCMC analyses are performed under the same model, starting from overdispersed values. Each panel present the time series plot of a (marginal) quantity along the MCMC: **A** the log-likelihood; **B** parameter of the gamma distribution modelling among-site rate variation (shape parameter α); **C** transition to transversion rate ratio (κ). Note that stationarity is reached at about the 500th step (vertical dash red line), from which sampling for inference can start (possibly with thinning – see text).

Equation (II.4) can be further simplified for the ED prior for rates since $p(R|T) = p(R)$. As one of the objectives of this section is to examine the effect of the molecular clock assumption on time estimates, σ^2 is not integrated out, but is estimated from the data; see below.

In general, it is very expensive to calculate the normalising constant $p(X)$ or the integral (II.4). Markov chain Monte Carlo (MCMC) (e.g. Gilks et al., 1996) is employed to approximate these (marginal) posterior distributions. At each step of the chain, a new state $\theta^* = \{R^*, T^*, \psi^*\}$ is proposed to change each parameter from a proposal distribution Q, which is assumed to be a normal distribution with the mean centred at the current state θ. The variance of the normal distribution is a tuning parameter. The new state θ^* is accepted with probability h (Metropolis et al., 1953; Hastings, 1970).

$$h = \min\left\{1, \frac{p(\theta^* \mid X)\, Q(\theta^* \to \theta)}{p(\theta \mid X)\, Q(\theta \to \theta^*)}\right\} = \min\left\{1, \frac{p(\theta^* \mid X)}{p(\theta \mid X)}\right\} \qquad \text{(II.5)}$$

The simplification is because the proposal distribution Q is symmetrical. The ratio of posterior distributions is still somewhat complicated. It involves the product of two terms: the likelihood ratio and the prior ratio, as shown in equation (II.5).

$$h = \min\left\{1, \frac{p(X \mid R^*, T^*, \psi^*)}{p(X \mid R, T, \psi)} \times \frac{p(R^* \mid T^*)\; p(T^*)\; p(\psi^*)}{p(R \mid T)\; p(T)\; p(\psi)}\right\} \qquad \text{(II.6)}$$

The updating scheme consists of two steps. In step 1, a divergence time is chosen at random to be updated together with parameters λ, μ, ρ of the BDP. In step 2, a rate for branch is chosen at random for updating. The tuning parameters for rates and times were adjusted by running preliminary chains to attain a balance between acceptance rate and mixing. If a proposed state, e.g. a node time and its hyperparameters, is not accepted, the algorithm moves to the next step.

Sampling from the posterior distribution can start when the chain has reached stationarity. The posterior probability $p(R, T \mid X)$ was found to reach stationarity quickly, whereas times and rates typically converged more slowly, especially for large data sets. Although there exist heuristic tests to know when the MCMC has converged, none of them seems infallible (Gilks et al., 1996). I have monitored convergence by plotting times series of the studied variables (times and rates). Four chains were run from very different starting points (see Figure II.4). Linear regressions were performed on times series of each variable, testing the significance of the slope: the *p*-value should be large, indicating a slope not significantly different from zero, and the autocorrelation functions should not detect any structure in the samples. Sampling starts after a burn-in period defined as the time the chain takes to forget the initial state and reach stationarity. The chain is sampled every 100 accepted states, hereby "thinning" the chain (Raftery and Lewis, 1996) and reducing autocorrelation between successive samples. I have used the median of the estimated posterior distribution as the best point estimate of that parameter. The following section shows this is because the median is less biased than the mean.

II.2 – Model selection and comparison with local clocks: two applications

II.2.a – Comparison of the different models of rate change

I analysed a small data set that consists of the tRNA-coding genes of the mitochondrial genome of six hominoid species: common chimpanzee, pygmy chimpanzee, human, gorilla, orang-utan and siamang (Horai et al., 1992). Alignment gaps were removed, leaving 762 nucleotides in the sequences. The phylogenetic relationship of these species seems well established and the tree shown in Figure II.5 will be assumed throughout. The data set was analysed under the HKY85 + Γ model of nucleotide substitution (Hasegawa et al., 1985; Yang, 1994b). The orang-utan divergence was set at 13 million years ago

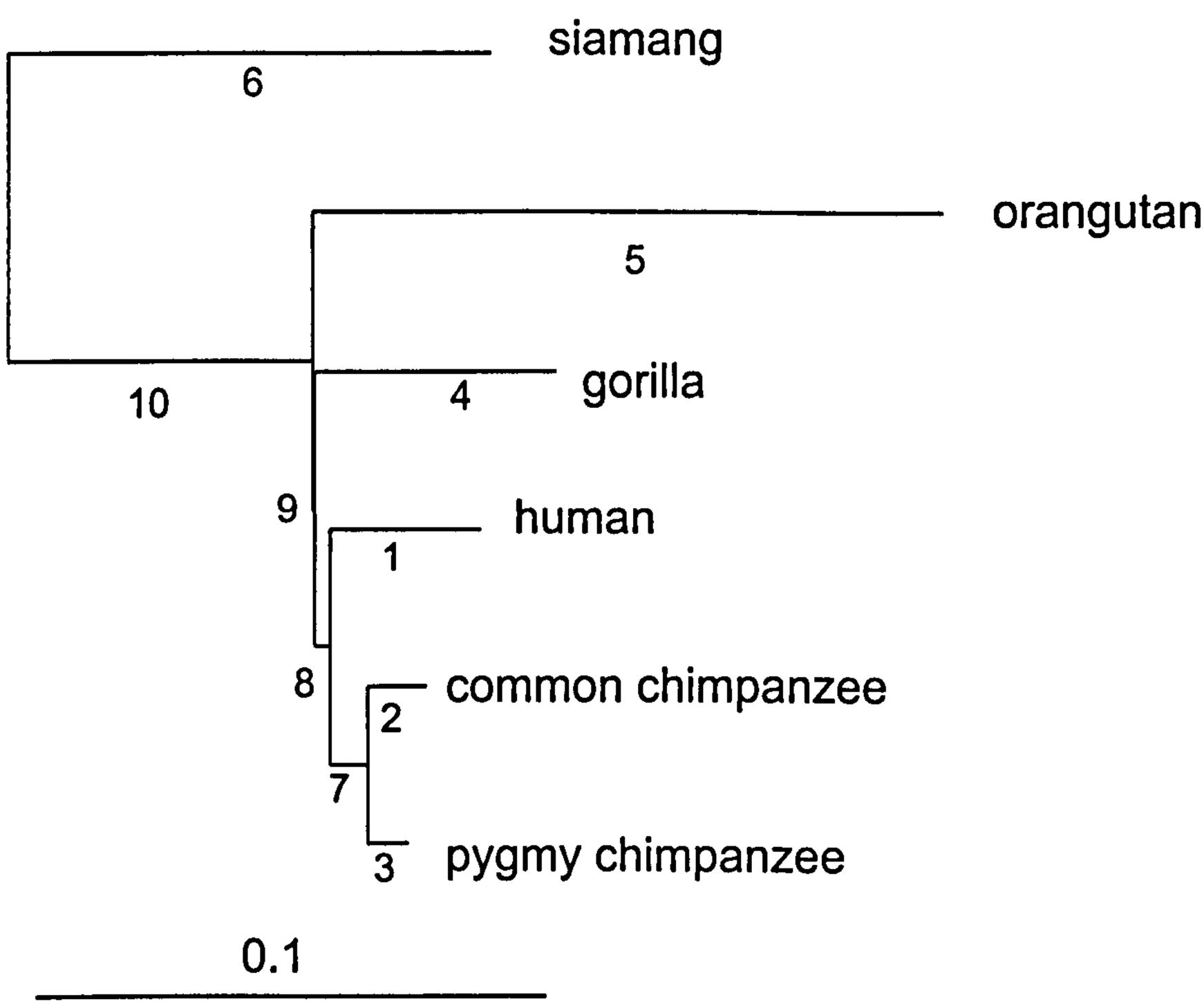


Figure II.5. ML tree for six species of hominoids. The branch lengths of the unrooted tree were estimated under the HKY85 + Γ model of nucleotide substitution. The root of the tree is placed on the siamang branch.

(MYA) and used as a calibration point (Horai et al., 1992). The molecular clock assumption was not rejected by the likelihood ratio test; the LRT statistic is $2\Delta\ell = 2 \times (-1785.96 - (-1789.65)) = 7.38$, with $P = 0.12$ and d.f. = 4. MLEs of substitution parameters without the clock are $\hat{\kappa} = 45.20$ and $\hat{\alpha} = 0.187$. These values were used in the MCMC runs in the Bayes analysis. Each chain was run with a burn-in of 10^4 steps, after which 10^4 samples were collected every 100 steps.

When the variance for the rate (σ^2) is very small, all the models essentially make the clock assumption, and produced similar estimates for the divergence times (Table II.1A). SLD has the largest L_k^{post}, but *PBF* is always less than 0.53 when this model is compared with any other model, so that the differences are not significant (see Kass and Raftery, 1995, p.777).

The exponential model does not have any hyperparameter to control its variance. In all other models, increasing the variance for the rate (σ^2) relaxes the clock assumption. Note that the same σ^2 in the different models means different extent of rate variation. Figure II.6 shows the influence of σ^2 on the estimates of two rates: r_5 for the branch ancestral to orang-utan and r_7 for the branch ancestral to the two chimpanzee species (see Figure II.5). It is clear that rates and divergence times are sensitive to the hyperparameter σ^2 for all models when σ^2 is small. When σ^2 is large, LND (not shown) and GD reach a plateau and the estimates are not sensitive to σ^2.

The probability L_k^{post} is maximised to estimate the hyperparameter σ^2 in the LND, SLD, and GD models, and β and σ^2 in the OUP model. Figure II.7A shows that the models behave differently. Under LND and GD, L_k^{post} reaches a plateau for large values of σ^2, and does not decrease until σ^2 is very large (results not shown). As discussed above, date and rate estimates are insensitive to σ^2 in these two distributions (Figure II.6).

Table II.1. Bayes estimates (posterior medians ± SE) of the divergence times in clock-like (A) and non-clock like (B) analyses.

	Chimpanzees	Human	Gorilla	Siamang	L_k^{post}
A – Clock-like analysis					
Clock	2.14 ± 0.70	4.79 ± 1.09	7.03 ± 1.38	18.33 ± 2.01	−1792.34
LND ($\sigma^2 = 10^{-4}$)	2.09 ± 0.65	4.69 ± 1.01	6.89 ± 1.29	19.03 ± 1.98	−1791.95
SLD ($\sigma^2 = 10^{-4}$)	2.08 ± 0.65	4.71 ± 1.02	6.92 ± 1.26	18.87 ± 1.95	−1791.90
GD ($\sigma^2 = 10^{-4}$)	2.11 ± 0.81	4.76 ± 1.21	7.00 ± 1.46	19.40 ± 2.12	−1792.43
OUP ($\beta = 10^2$, $\sigma^2 = 10^{-4}$)	2.17 ± 0.87	4.79 ± 1.28	7.02 ± 1.54	19.22 ± 2.16	−1792.29
B – Non clock-like analysis*					
LND ($\sigma^2 = 10$)	5.66 ± 2.63	8.82 ± 2.63	10.96 ± 2.60	17.07 ± 2.50	−1790.17
SLD ($\sigma^2 = 1$)	5.51 ± 1.92	8.64 ± 2.04	10.49 ± 2.10	16.57 ± 2.14	−1790.73
GD ($\sigma^2 = 9$)	5.02 ± 2.84	7.99 ± 2.82	10.38 ± 2.77	16.92 ± 2.58	−1790.08
OUP ($\beta = 10^2$, $\sigma^2 = 1$)	4.54 ± 1.95	7.89 ± 2.08	10.21 ± 2.09	15.09 ± 1.79	−1788.75
ED	5.89 ± 2.21	9.11 ± 2.20	11.02 ± 2.13	15.01 ± 1.91	−1789.72

*Hyperparameters β and σ^2 are chosen to maximise L_k^{post}.

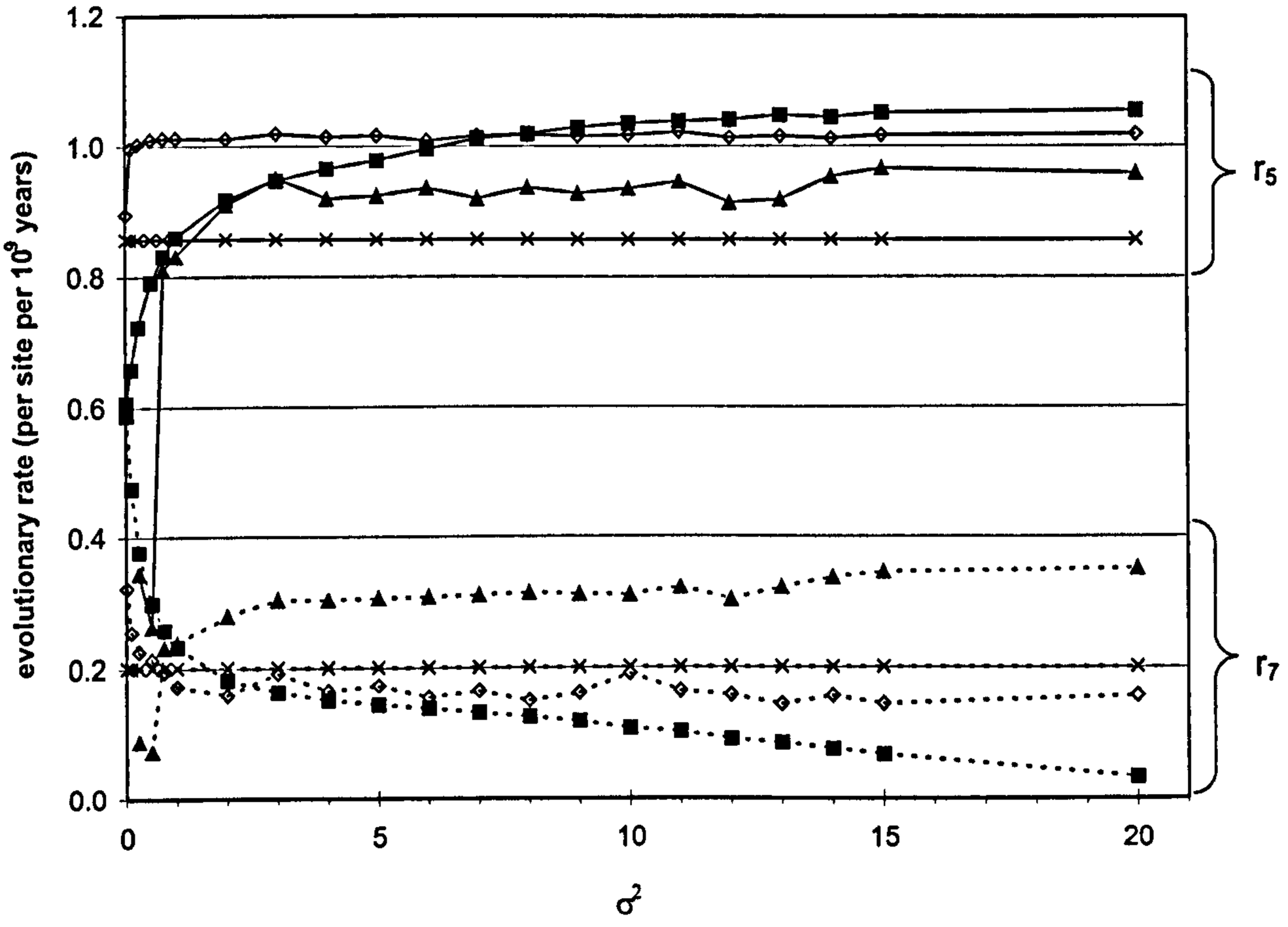


Figure II.6. Posterior medians of evolutionary rates for branches 5 and 7 in Figure II.5 under different models of rate change: SLD (■), OUP (▲), GD (◊) and ED (×). Rates are measured by the expected number of substitutions per site per 10^9 years. The hyperparameter β of OUP is set to 100.

Under SLD and OUP, L_k^{post} is sensitive to σ^2. The optimum value under SLD is about $\sigma^2 = 1$ (Figure II.7A), when the optimum values for OUP are about β=100 and $\sigma^2 = 1$ (Figure II.7B). As discussed above, estimates of dates and rates are somewhat sensitive to these hyperparameters under SLD and OUP, and the optimum values of σ^2 and β were used in Table II.1. Note that when optimum parameters are used, the date estimates are similar among the different models.

I also used L_k^{post} to compare models of rate change (Figure II.7A). For OUP, the hyperparameter β has been set to 100, which is close to the optimal value. OUP outperformed the other models. The *PBF*, computed from those probabilities, ranges from 1.0 to 2.0 on the log scale for comparison between OUP and the other models (Table II.1B), indicating a small preference for OUP (Kass and Raftery, 1995).

II.2.b – Comparison with ML analysis under local clock models

Comparison of the Bayesian approach (Table 1) with maximum likelihood (Table 2) is a good means of testing the MCMC implementation. Under the molecular clock assumption, both approaches should give similar estimates, with larger SEs from the Bayes models. We note that the ML date estimates for nodes younger than the calibration point are slightly younger (say 4.3 MYA for the human-chimpanzee divergence) than the Bayes estimates (4.7 to 4.8 MYA). For nodes older than the calibration point, the difference is also small but in the opposite direction. The observed discrepancy appears to be due to the BDP prior for divergence times used in the Bayes approach. The use of a small sampling fraction, $\rho \sim U(0, 0.01)$, has the effect of shortening the internal branches (Yang and Rannala, 1997).

When the molecular clock is relaxed, the MLEs of dates are very different from and much older than those under the clock (Table 2). For example, the date for human-

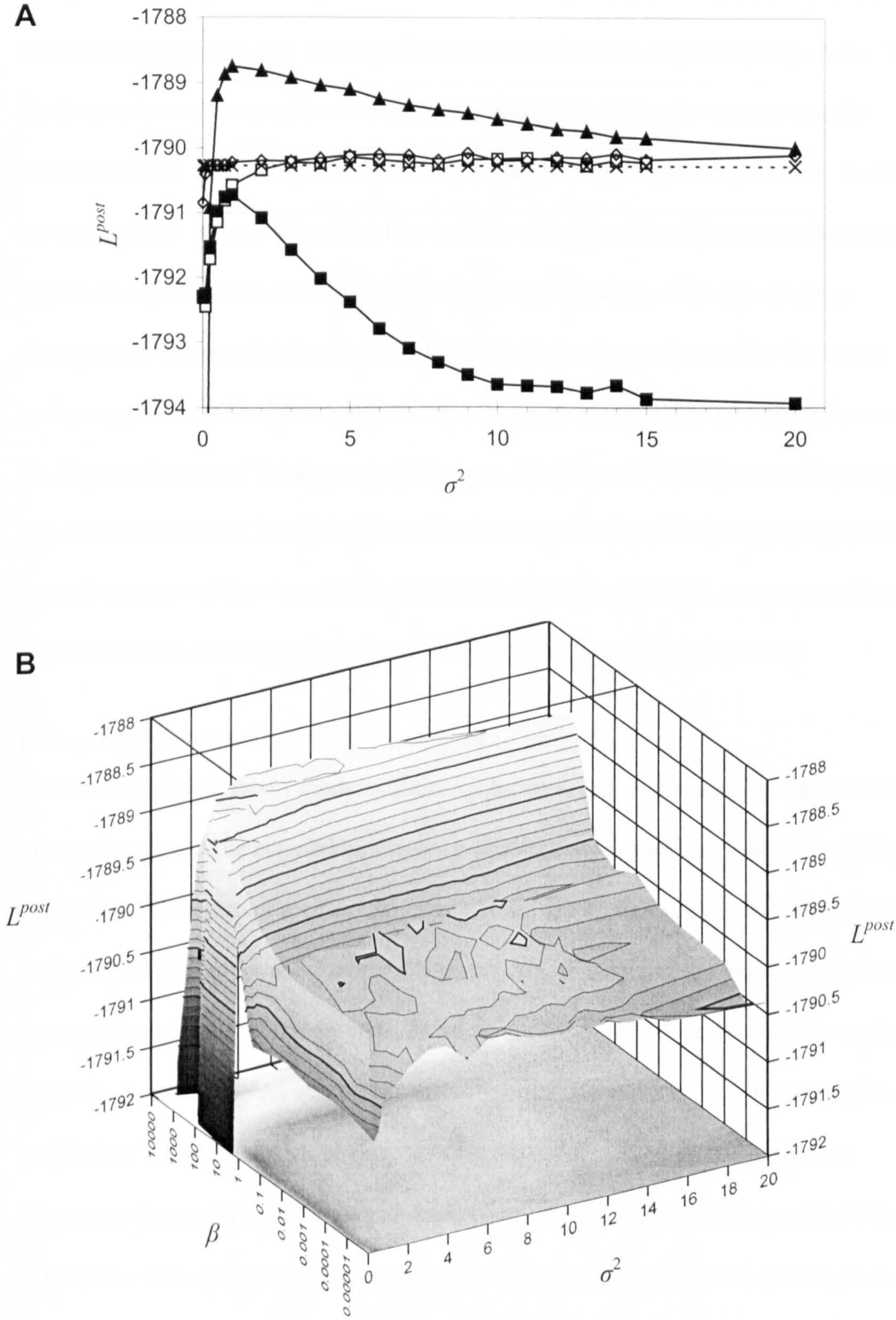


Figure II.7. **A** Approximate L_k^{post} as a function of the hyperparameter σ^2 under different models of rate change: LND (□), SLD (■), OUP (▲), GD (◊) and ED (×). Under OUP, $\beta = 100$ is fixed. **B** Approximate surface of L_k^{post} under OUP as a function of the hyperparameters β and σ^2.

chimpanzee divergence changed from 4.3 MYA under the clock to 8.8 MYA under a two-rate local clock model, although both estimates involve large sampling errors. Thus the date estimates are sensitive to the clock assumption, although we note that the molecular clock was not rejected by the LRT. In the Bayes approach, relaxing the clock assumption also had considerable effect on date estimates. The chimpanzee, human and gorilla divergences become much older than under the clock, while the siamang divergence becomes slightly younger. While the Bayes date estimates for the recent nodes are similar to the MLEs, either under the clock or when this assumption is relaxed, the Bayes estimates of the siamang divergence date (15 to 17 MYA) are much younger than the MLEs (about 36 to 58 MYA). The last estimate (58 MYA) is probably too old, but a divergence of the siamang (*Hylobatidae*) as early as 16 MYA, as suggested by the Bayes estimates, would imply a very rapid diversification of the *Hominidae*.

II.2.c – Application to the 18S rRNA data set

To test the different Bayes models of rate evolution and demonstrate the important effect of rate change on date estimation, I have re-analysed the nuclear-encoded 18S rRNA genes from 39 metazoan species (Bromham et al., 1998), rooted by a fern, Polypodium. As reviewed by Cooper and Fortey (1998), the time of origin of the animal phyla has been controversial. A common view, based on the fossil records, holds that the early Cambrian (ca. 545 MYA) was characterised by an accelerated evolution marking an "explosion" of the metazoan phyla (e.g. Valentine et al., 1996). In particular, the divergence between protostomes and deuterostomes is thought to have occurred about 600 MYA. However, molecular studies such as that of Bromham et al. (1998) produced estimates as far back as about 1,200 MYA, almost twice as old.

The sequences consist of 1,710 nucleotides. Gaps were removed from the alignment, and the data set was analysed under the HKY85 + Γ model of nucleotide substitution,

Table II.2. Maximum likelihood estimates of divergence times (± SE) under the clock and local-clock models.

	clock (1 rate)	2 rates	3 rates	4 rates
Chimpanzees	1.77 ± 0.54	3.74 ± 1.13	3.76 ± 1.16	5.88 ± 1.82
Human	4.28 ± 0.91	8.85 ± 1.86	7.59 ± 2.61	10.68 ± 3.80
Gorilla	6.50 ± 1.18	13.00 ± 2.32	13.00 ± 2.54	12.64 ± 6.13
Orang-utan	13	13	13	13
Siamang	19.56 ± 3.49	35.86 ± 6.35	37.61 ± 7.08	57.57 ± 11.51
ℓ	-1773.21	-1770.90	-1770.52	-1769.42
$\hat{r}_1$	1	3.41	3.60	6.23
$\hat{r}_2$	1	1	1.51	1.84
$\hat{r}_3$	1	1	1	2.45

ML analyses were performed under the HKY85 + Γ model (κ= 45.20, α= 0.187). The calibration point was set at 13 MYA for the orang-utan. Local clock settings: r_1 for orang-utan; r_2 for human; r_3 for gorilla, and r_0 = 1 for all other branches.

with the transition to transversion rate ratio and the shape parameter of the Γ distribution set to their MLEs obtained without the clock ($\hat{\kappa} = 3.461$ and $\hat{\alpha} = 0.373$). The tree topology was fixed (Figure II.8), according to Nielsen (1995). The molecular clock assumption was rejected by the LRT; the test statistic is $2\Delta\ell = 2 \times (-13{,}948.10 - (-14{,}381.85)) = 867.50$, $P < 0.01$. The shape of the ML tree under no clock (not shown) indicated very variable rates among lineages, which may preclude traditional analyses, either by ML local clocks (Yoder and Yang, 2000) or by linearizing the tree (Takezaki et al., 1995).

The Bayes analysis was conducted by drawing the hyperparameters of the BDP prior for times from uniform distributions, $\lambda \sim U(0, 15)$, $\mu \sim U(0, 5)$, and $\rho \sim U(0, 0.001)$. I used the ED, SLN, and OUP models of rate change. MCMC runs included a burn-in period of 10^5 steps, after which 10^5 samples were collected every 100 accepted states. I averaged the posterior estimates over eight calibration points given by Bromham et al., 1998): Collembola-Pterygota, 390 MYA [1]; Aranaea-Scorpionida, 405 MYA [2]; Arachnida-Merostomata, 520 MYA [7]; Cephalochordata-Chordata, 530 MYA [8]; Coelacanth-Dipnoi/Tetrapoda, 418 MYA [3]; Osteichthyes-Dipnoi/Tetrapoda, 428 MYA [4]; Agnata-Gnathostoma, 510 MYA [6]; Asteroidea-Echinoidea, 500 MYA [5] (numbers in square brackets refer to Figure II.8).

Dates estimated with a Bayesian clock-like model, with a small variance for the prior on the rates, place the echinoderms-chordates and protostomes-deuterostomes divergences at 1,205 MYA (95% credible set 1,062-1,341) and 1,450 MYA (95% credible set 1,321-1,567) respectively. These estimates are very similar to the ones found by the original authors (compare with Figure 2 of Bromham et al., 1998). Our date estimates are nonetheless smaller. This bias can be explained by the parameterisation of the BDP (see below).

Table II.3. Fit of different models of rate change to the metazoan 18S rRNA data set.

Model	β	σ^2	L_k^{post}
Clock	n.a.	n.a.	-14,418.07
ED	n.a.	n.a.	-14261.44
SLD	n.a.	1	-14353.29
	n.a.	10	-14147.97
	n.a.	20	-14340.55
	n.a.	40	-14286.99
OUP	0.01	1	-14362.74
		10	-14437.74
		20	-14000.97
		40	-13991.73
	0.1	1	-13988.96
		10	-13988.83
		20	-14012.19
		40	-13988.93
	1	1	-13988.05
		10	-13986.44
		20	-14004.27
		40	-13992.62

n.a.: not applicable.

To relax the molecular clock hypothesis, the ED, SLN, and OUP models of rate change have been evaluated with an empirical Bayes phase to estimate the hyperparameters. The results are summarised in Table II.3. The *PBF* given each model and parameterisation is provided only for the values around the maximum of L_k^{post}, although an extensive search has been carried out to make sure there were no other optima. Again, OUP explains the data better than any other model. The maximum of L_k^{post} is around $\beta = 1$ and $\sigma^2 = 10$, but the probability surface in this region is almost flat (Table II.3). The estimated σ^2 for SLN is around 10. This large value is consistent with the large statistic in the LRT of the clock, and indicates that rates are more variable than in the small hominoid data set.

The estimates of the divergence times under the ED model, summarised in Figure II.8, are very similar with those under SLN and OUP (not shown). The time estimates are consistent with the fossil records (e.g. Conway-Morris, 1998a), or with linearized analyses performed on many genes (Ayala et al., 1998). This latter analysis, of 18 protein-coding genes, estimated divergence dates at 628 ± 76 MYA for the echinoderms-chordates split and at 736 ± 65 MYA for the protostomes-deuterostomes separation. Our estimates, for a single gene, are respectively 550 MYA (95% credible set: 510-574) and 560 MYA (95% credible set: 522-581) under ED, and 579 MYA (498-608) and 595 MYA (519-616) under OUP ($\beta = 0.1$ and $\sigma^2 = 10$). As there is no need to eliminate outlying taxa, all the available information in the gene is taken into account in the Bayesian approach.

Possible biases must be considered when interpreting the results of the Bayesian analysis. Firstly, I used a fixed tree topology, while uncertainty exists regarding the evolutionary history of the metazoan phyla. The effect of the uncertain phylogeny on date estimation deserves consideration, although a pervious study (Yoder and Yang, 2000) suggested that

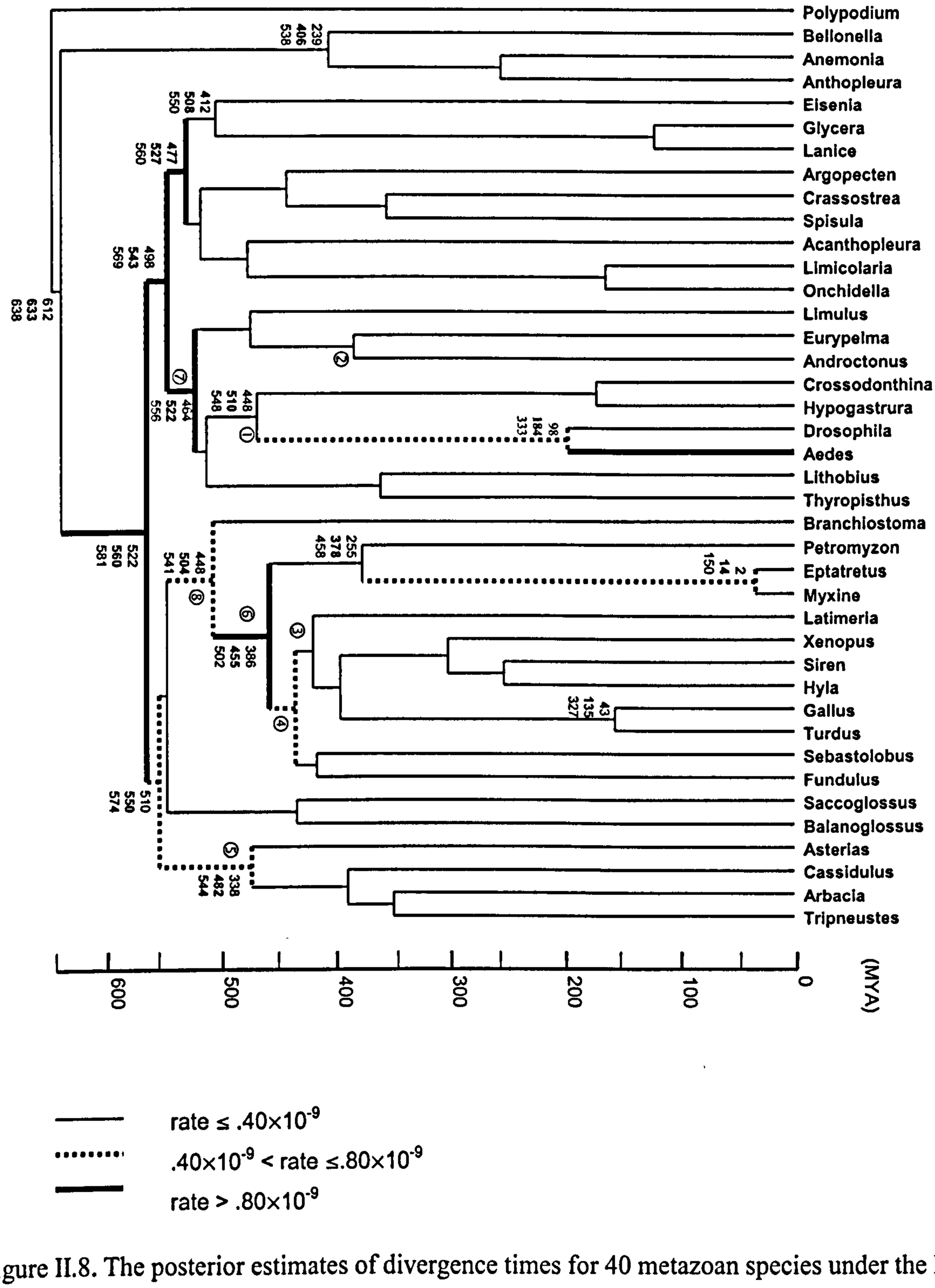


Figure II.8. The posterior estimates of divergence times for 40 metazoan species under the ED model of rate change. Calibration points from fossil dates (see main text) are indicated by circled numbers. Estimates were obtained under the HKY85 + Γ model of nucleotide substitution. Branch lengths are scaled to time, and the thickness of a branch indicates the evolutionary rate (expected number of substitutions per site per 10^9 years). Numbers at internal nodes represent, from top to bottom, the lower limit of the 95% credible set, the time estimate (bold) and the upper limit of the 95% credible set.

plausible topologies gave similar speciation date estimates. Secondly, I have found in the hominoid data set that Bayesian inference may be sensitive to the hyperparameter σ^2 of the prior model of rate change. Similar effects were found in the metazoan data set. For example, date estimates under the clock assumption were drastically different from those presented in Figure II.8. In this regard, I note that estimates of dates under different models were similar when optimum values of σ^2 were used in each model of rate change. Lastly, the results presented here are obtained from one single gene, and should be taken with caution.

The rate estimates (Figure II.8) suggest that the metazoan 18S gene underwent a complex history, with high evolutionary rates during the Cambrian (between 550 and 500 MYA) for triploblastic animals, while diploblastic animals had much lower rates, which they seem to have conserved to date. The episode of high evolutionary rate in the Cambrian was followed by a steep decline to a more or less steady rate for protostomes, whereas the pro-chordates underwent another burst at approximately late Ordovician / Silurian. Subsequent rate accelerations were detected for the branches leading to the Myxiniformes and the Diptera, with a burst for the Nematocera. The history of the 18S rRNA gene might therefore not be characterised as a mere decline of rates as suggested recently (see Bromham and Hendy, 2000), although the reasons for this "episodic evolution" (Gillespie, 1991) are not yet understood.

II.2.d – Conclusions

Analyses of both the hominoid and metazoan data sets suggest that date estimates are very sensitive to the molecular clock hypothesis. It should be noted that most molecular dates have been based on the simplifying assumption of the molecular clock, although some methods have been proposed to constrain a data set to conform to this hypothesis,

for example by tree linearization (Takezaki et al., 1995). Bromham et al. (2000) pointed out that the power to detect rate variation might not be very high, and as a result, use of such tests to filter data might still lead to systematically biased date estimates. As demonstrated by our analysis of the metazoan data set, the Bayes approach offers a promising alternative to the problem, estimating divergence dates while detecting and accommodating possible rate variation.

The likelihood-based local clock models (Yoder and Yang, 2000) are useful if prior information is available about which lineages might have different rates. For instance, these models are useful for testing whether certain groups of species, such as primates vs. rodents, have different evolutionary rates. When such information is unavailable, it is more natural to resort to a Bayes model of random change, although at a greater computational cost. In contrast to other implementations (e.g. Thorne et al., 1998) where the variance of the model of rate change is drawn from an informative exponential distribution, the empirical Bayes approach is a possible way to estimate the departure from the clock assumption. While the approach appears most appropriate when rates change slowly over time or branches, it can accommodate rapid rate changes with the use of large values of σ^2, as shown by the analysis of the metazoan data set. Our results suggest that beyond a certain value, the hyperparameter σ^2 has little influence on the posterior mean of the target distribution.

The use of the posterior Bayes factor appears contentious (see discussion to Aitkin, 1991), but is here the most operational selection procedure. Our approach of estimating the hyperparameter σ^2 (or β and σ^2 in OUP) does not properly account for the uncertainty concerning those hyperparameters, as the optimum values were treated as known when divergence dates were estimated. A full Bayes approach should integrate over β and σ^2 (see Chapter III). I attempted to estimate them, applying such an approach to both data

sets analysed in this section, averaging over uniform priors for β and σ^2 in the MCMC. However, I found that the chain did not converge well, in particular regarding the marginal distributions of β and σ^2. As discussed above, the probability surface was relatively flat for large values of σ^2. It is possible that simultaneous use of multiple calibration points might provide information about rates and thus help with the convergence of the MCMC in a full Bayes analysis.

The general approach to allow rates to vary in time appears promising. The metazoan divergence date estimates are closer to what is expected from palaeotonlogical data (see also Chapter III). Similar results follow from the analysis of the total mitochondrial genome of six mammals (see Appendix 1), where the estimated divergence dates between mouse / rat and primates / rodents are closer to conventional wisdom. However, this general approach of allowing rates to vary does not give satisfactory results for the small primate data set. Both local molecular clocks and Bayesian models of rate change give date estimates of the human / chimpanzee divergence about twice as old as those generally accepted.

Apart from the small primate data set, the general trend introduced by allowing rates to vary in time seems to be that the estimated dates tend to be closer to the present. This is what is expected when early lineages evolve with high rates, as found by Bromham and Hendy (2000). It would therefore be interesting to test the approach on a group of lineages known or expected to show rate acceleration.

II.3 – Sensitivity analysis

A number of models of rate change have been presented above to relax the molecular assumption, but their merit has only been assessed by their fit to the data as measured

by *PBF*. Here, simulations were carried out to evaluate their properties under different hypotheses, in particular when data do not conform hypotheses either relative to the model of rate change, or to the speciation model. It is shown that estimation of divergence times is affected by departures from non-correlated rate change patterns. Improvements by means of constraints on the node times are discussed.

II.3.a – Model and simulation conditions

The general Bayesian model

In order to relax the molecular clock assumption and to obtain better estimates of divergence times, different Bayesian models of rate change were formulated above (see section II.1.c). These models have the common feature of modelling rate change using a continuous distribution or process, and setting a prior for divergence times. From the Bayes theorem, the posterior distribution of the vector of parameters is proportional to the likelihood times the joint distribution of rates and times. In particular, the marginal posterior probability of the vector of divergence times given the sequence data X is derived as:

$$p(T) = \frac{\iint p(X \mid R,T,\theta)\, p(R,T)\, d\theta\, dR}{\iiint p(X \mid R,T,\theta)\, p(R,T)\, d\theta\, dR\, dT} \qquad \text{(II.7)}$$

where $p(X \mid R, T, \theta)$ is the traditional likelihood (Felsenstein, 1981) of a set of parameters: the rates of evolution (R), the divergence times (T) and parameters of the substitution model (θ). The joint probability $p(R, T)$ represents our prior belief about the processes thought to have generated the observed data X. The denominator, representing the probability of the data, is a constant and will be hereafter denoted $p(X)$. Equation (II.7) assumes that the topology is fixed (see Chapter V for a relaxation of this hypothesis). Under the simple JC69 substitution model (Jukes and Cantor, 1969) used here, equation (II.7) reduces to:

$$p(T) = \int p(X \mid R,T)\, p(R,T)\, dR \Big/ p(X) \qquad \text{(II.8)}$$

For its simplicity and because only rates and times affect the likelihood, this model was chosen to investigate the performance of different models of rate change under different simulation conditions.

Prior distributions for times and rates

Divergence times were assumed to follow one of these two models: (i) a generalised birth and death process with species sampling ρ where the speciation rate λ and the extinction rate μ are assumed to be constant per lineage (Yang and Rannala, 1997; section II.1.b); (ii) a uniform process, corresponding to an uninformative prior. The rate of a branch can be assumed to follow different models (section II.1.c). Here, I focus on four of them: the Bayesian equivalent of the molecular clock, the stationary lognormal distribution (SLD), the exponential distribution (ED) and the Ornstein-Uhlenbeck process (OUP). Only the SLD and the OUP models have hyperparameters and can be noted as SLD(σ^2) and OUP(β; σ^2), where σ^2 is a variance term that describes the relaxation of the clock, and β is a friction term. The clock assumption is relaxed for large values of σ^2 and small values of β, as shown above (section II.1.c). However, unlike the model described above, these hyperparameters are integrated out here, assuming that σ^2 follows a gamma distribution of mean 15 and variance 25 and that β follows a lognormal distribution of mean log(.5) and variance .75. These prior distributions were chosen as being vague enough in the region of the parameter space where the clock is relaxed. The model of equation (II.8) becomes:

$$p(T) = \int_{\Xi} p(X \mid R,T)\, p(R \mid T,\beta,\sigma^2)\, p(T \mid \lambda,\mu,\rho)\, dR\, d\xi \Big/ p(X) \qquad \text{(II.9)}$$

where ξ is a vector of parameters (β, σ^2, λ, μ, ρ) on the space Ξ. The marginal distributions of equation (II.9) are approximated by a Markov chain Monte Carlo (e.g. Gilks et al., 1996) as described above (section II.1.e).

Simulation details

To study the properties of the models of rate change and the two speciation models, I considered a tree with eight ingroup taxa plus an outgroup. The four situations depicted in Figure II.9 were simulated: Tree (A) follows a strict molecular clock; Tree (B) simulates a constant slowdown of the rate of evolution correlated across all the ingroup taxa, where the rate is halved at each speciation event. This situation corresponds to the hypothesis underlying our models of rates change where rates are autocorrelated from ancestor to descendents. The last two situations, Tree (C) and (D) correspond to some further violations of the molecular clock assumption, where rates change in an uncorrelated manner (Tree (C)), or the tree exhibits a burst of evolution followed by an immediate slowdown of the rates in one group of taxa. This situation may correspond to some plausible scenarios (Chapter III). To sum up, Table II.4 presents the number of distinct values that are expected for the estimated branch lengths, rates and divergence times. Note that none of the simulations respect the assumptions of the speciation models, except Tree (A) with the uniform model. Under each condition, 100 replicates were simulated under the JC69 substitution model with sequences 1,000 nucleotides. All the simulations were done with the `evolver` program of the PAML package (Yang, 1997b). Data sets were analysed with a burn-in period of 20,000 states, after which chains were sampled for inference every 1,000 steps until a total of 500 states were collected for each replicate.

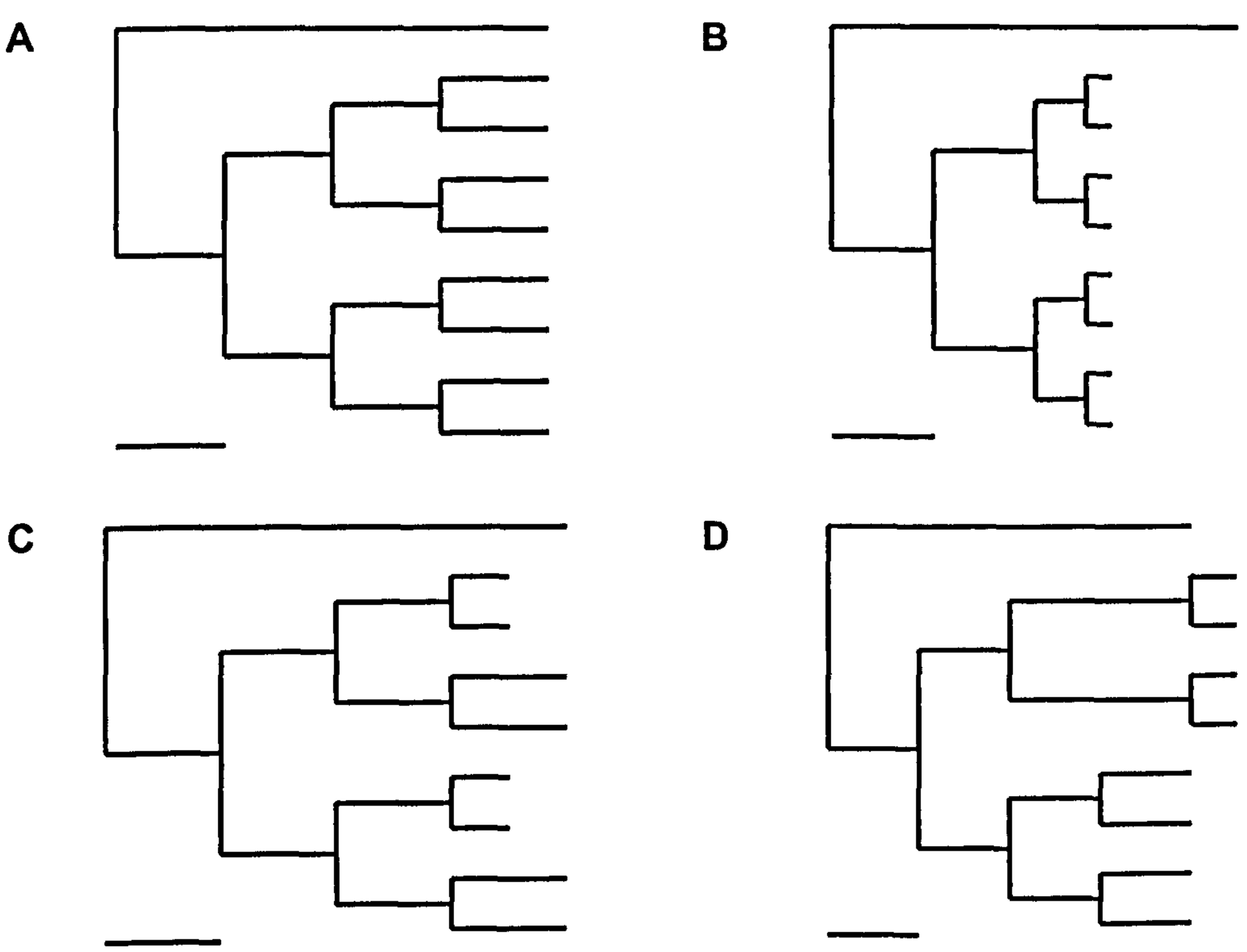


Figure II.9. The four trees used in the simulation study. Scale: the bar corresponds to 0.1 substitutions per site per unit of time. See text for details.

II.3.b – Performance of the models in simple cases

Simulation conditions under the molecular clock (A) and the correlated rate change (B) represent the simplest cases. In both cases, the estimated branch lengths are centred on the simulated "true" values, irrespective to the model of rate change used to analyse the data (Figure II.9). Under the uniform prior model describing speciation, the same pattern as in Figure II.10 is obtained (not shown). Table II.4 can be used to know how many distinct rate categories are expected to be recovered. In particular, simulations (A) should return one rate. Figure II.11 shows that while it is approximately the case under SLD and ED, OUP tends to be less flexible and overestimates rate variation. For all the models of rate change under simulation (A), the modal value of the posterior distribution is less than the rate estimate under the clock. This effect seems to be compensated by a much larger variance under the different models of rate change. In the case of simulation condition (B), three distinct categories of rates are expected. Both ED and SLD show posterior distributions centred on three modal values (Figure II.11). Each mode is approximately in a ratio of one half with the other, as simulated. However, OUP has much broader posterior distributions, and the three types of posterior distributions are largely overlapping. The effect of the model of speciation on the estimation of the rates is also clear from Figure II.11. Rates appear generally underestimated under models of rates change. This underestimation is more striking under a uniform prior for the divergence dates than under the BDP prior. When compared with the estimates under BDP, the divergence times estimated under the uniform prior distribution appear larger.

Examination of the estimates for divergence times under the uniform and the BDP priors suggest that they are better under BDP. For instance under the clock, the time expected from the simulation should be 0.25 units for the most recent nodes, which is approximately what is obtained under BDP whereas estimates are larger (older) under the

Table II.4. Summary of the simulation characteristics.

	(A)	(B)*	(C)	(D)
NBL	1	3	2	3
NR	1	3	2	3
NT	3	3	4	4

Notes – NBL: number of distinct branch lengths; NR: number of distinct rates; NT: number of distinct divergence times. An asterisk (*) indicates the simulations that fit the assumptions of the models of rate change.

uniform process (Figure II.12). This is in particular true under SLD; ED and OUP seem less sensitive to the prior distribution on divergence times. The reasons for these differences are not clear. Note also that the time estimates under SLD are generally larger than under ED, themselves greater than under OUP. This complicated process may lead to underestimates of divergence times on these simulated data. This is particularly clear under clock-like simulation conditions (tree A).

II.3.c – Robustness of the models

Real data are certainly more complicated than the two situations depicted by the scenarios (A) and (B). Although idealised situations are still simulated under (C) and (D), rates are no longer autocorrelated. This violates one assumption of the models of rate change. The effects are more difficult to interpret, but Table II.4 can be used again. While the three models of rate change seem to estimate the correct number of rate categories (Figure II.13), SLD does not distinguish the two most recent node times under simulations (C) and (D). However, the burst of evolution of simulation (D) is detected by the three models of rate change (Figure II.13 columns D). This suggests that these models are flexible enough to accommodate complex changes of rates of evolution across lineages.

When considering estimates of the divergence times, SLD does not distinguish two of the nodes. This follows what has been noted for the rates. The same biases are noted here as in the simpler cases: OUP tends to underestimates times, while a uniform process gives estimates larger than under BDP.

To conclude, it appears that OUP does not seem to be a very flexible process, at least as it is implemented here. One explanation is that the distributions chosen to integrate its hyperparameters out may be too narrow and still constrain the model too much. This could explain its poor performance with data simulated under the clock. OUP is also the

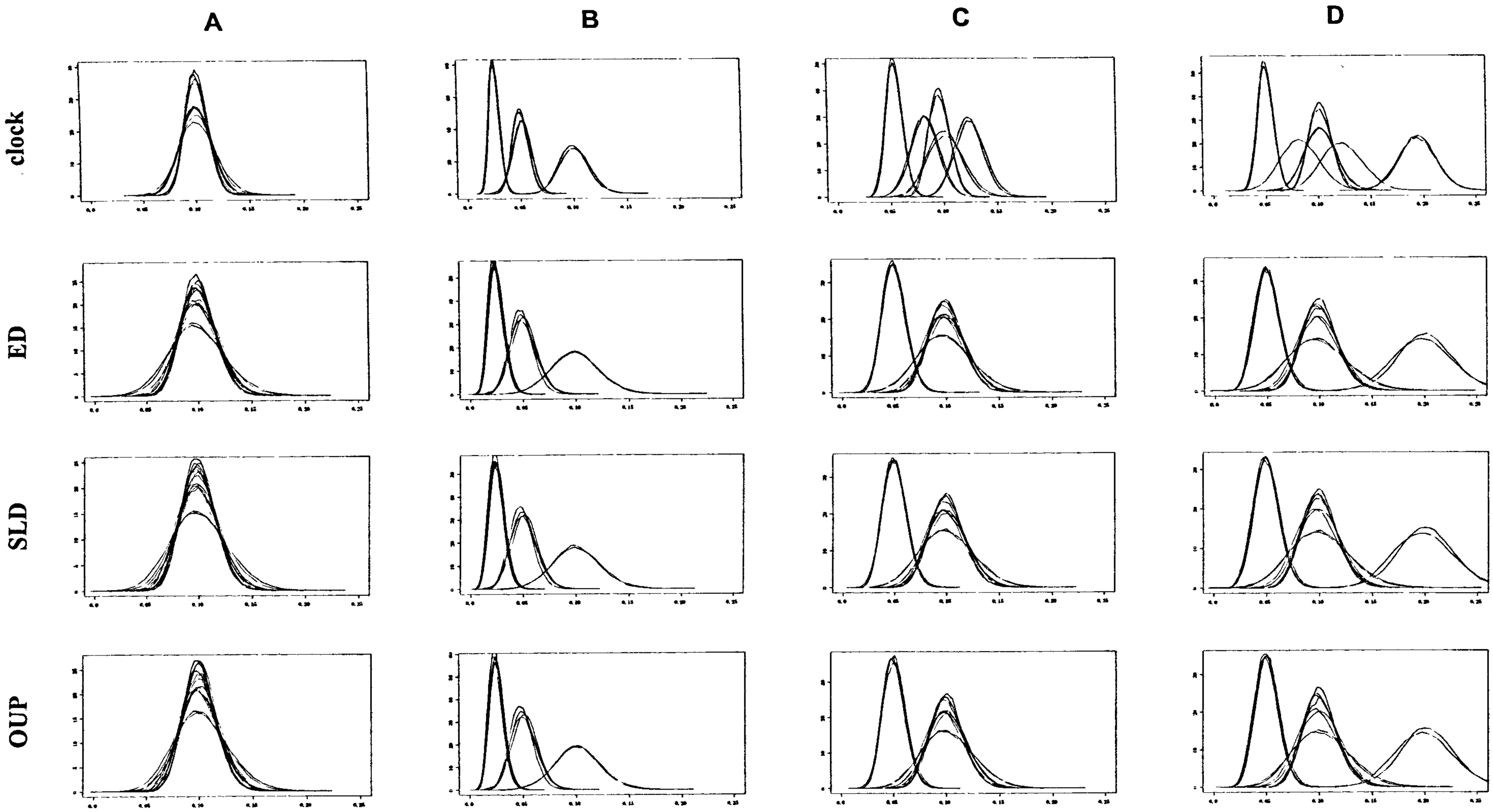


Figure II.10. Posterior distributions of the branch lengths under the four simulation trees. All the analyses are performed under the BDP model of speciation. The different models of rate change are: the clock (first line), the ED model of rate change, SLD and OUP (last line).

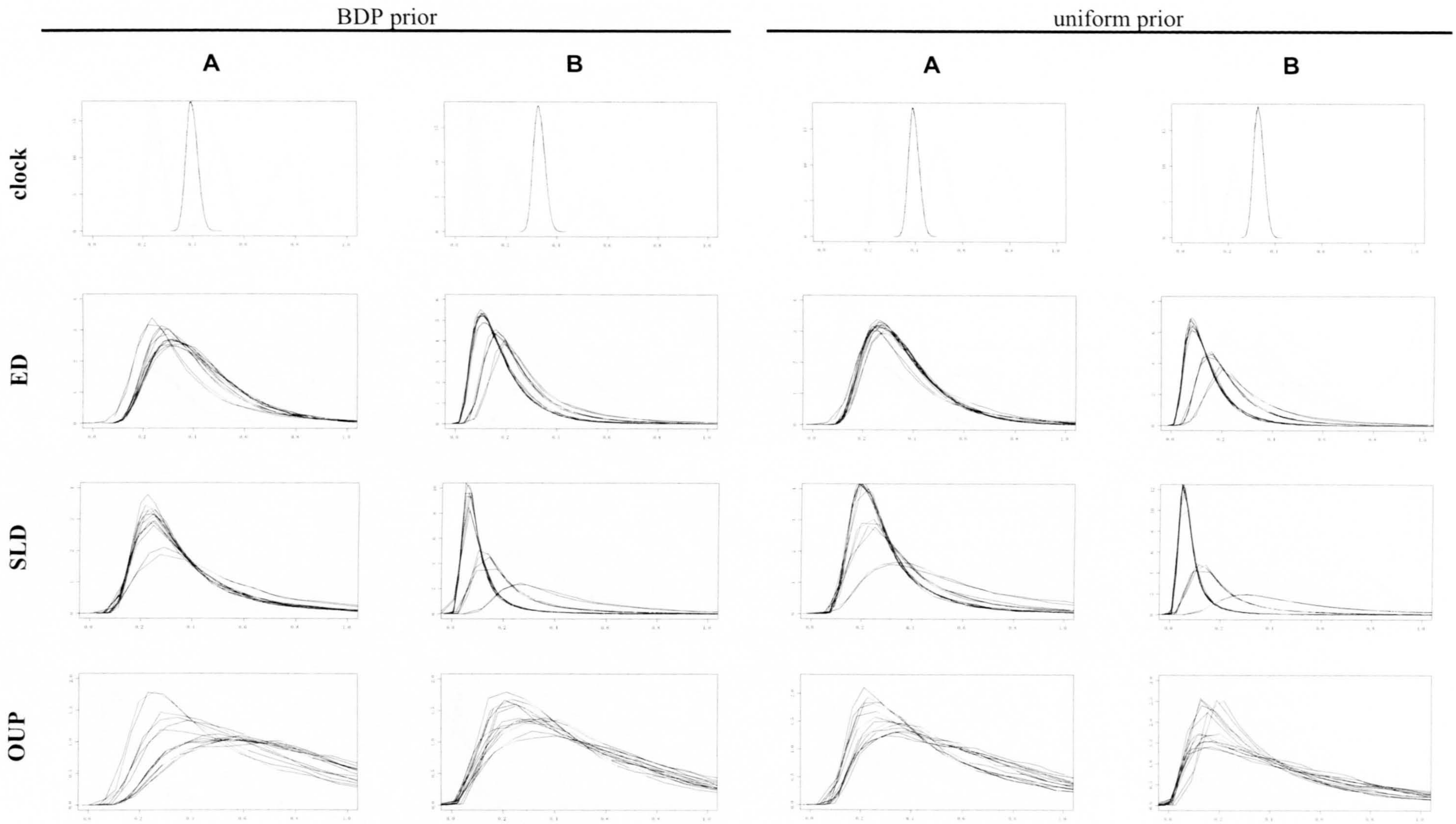


Figure II.11. Posterior distributions of the rates under the simple simulation trees. The different models of rate change are: the clock (first line), the ED model of rate change, SLD and OUP (last line).

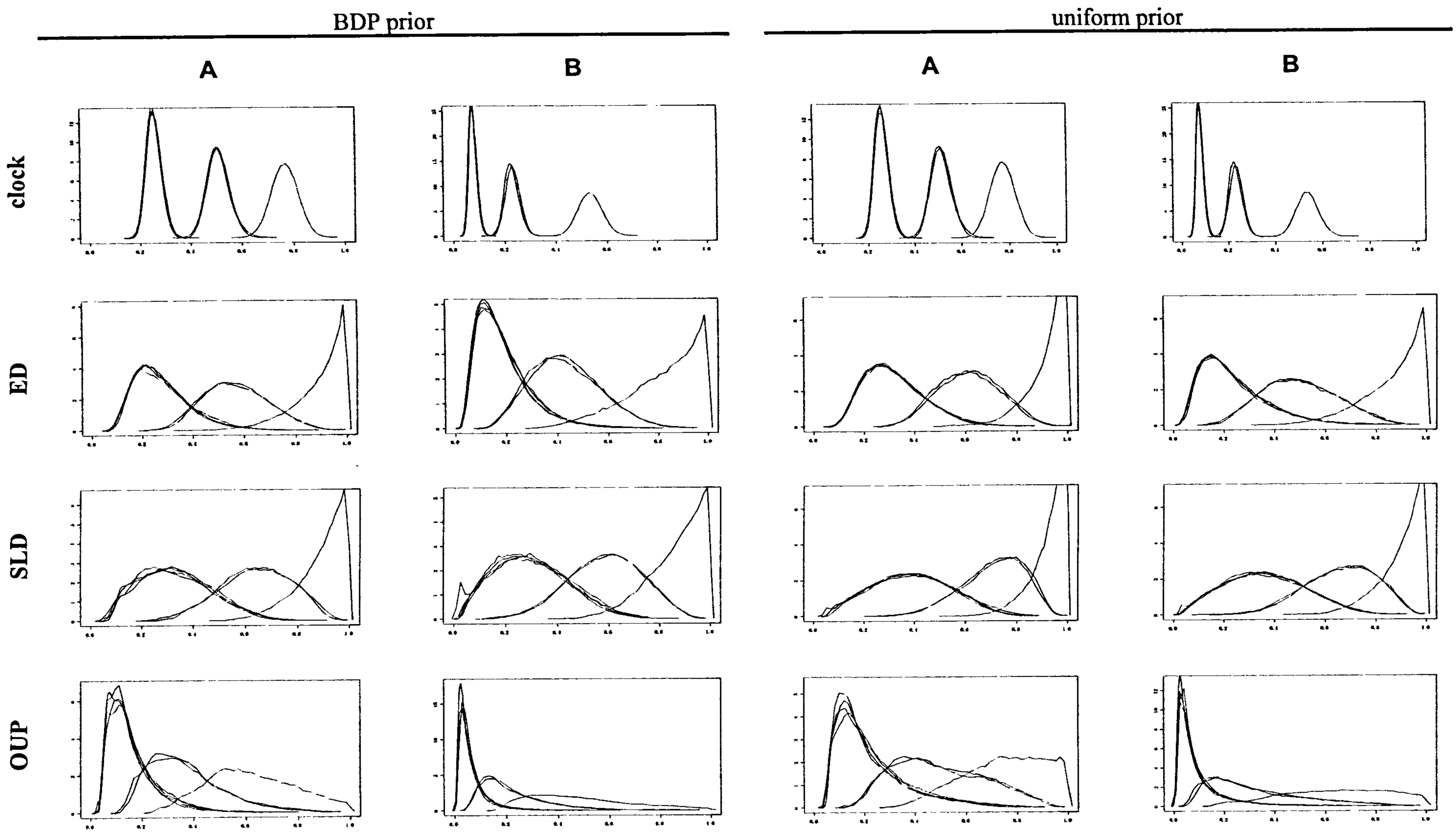


Figure II.12. Posterior distributions of the divergence times under the simple simulation trees. The different models of rate change are: the clock (first line), the ED model of rate change, SLD and OUP (last line).

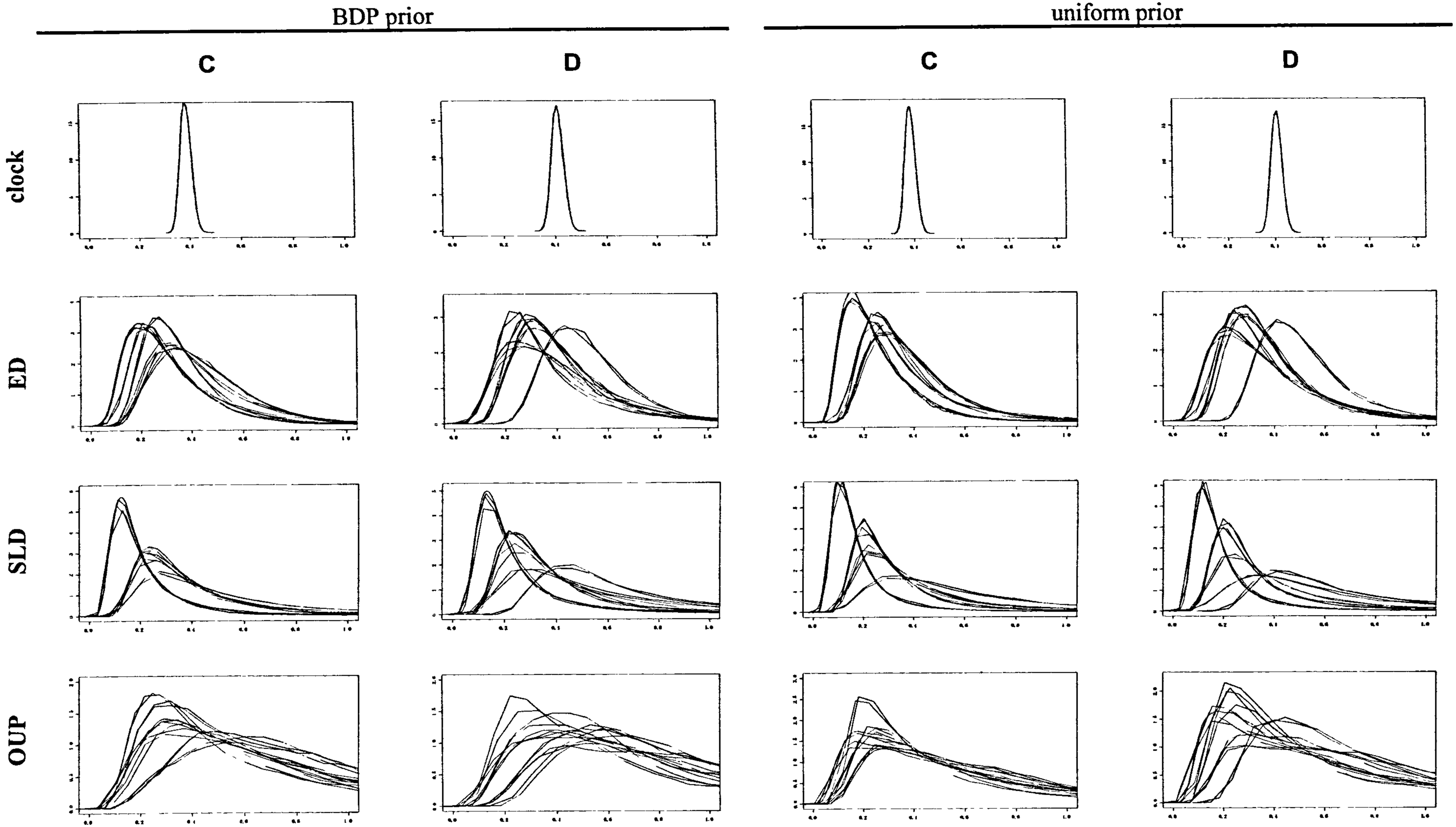


Figure II.13. Posterior distributions of the rates under the more complicated simulation trees. The different models of rate change are: the clock (first line), the ED model of rate change, SLD and OUP (last line).

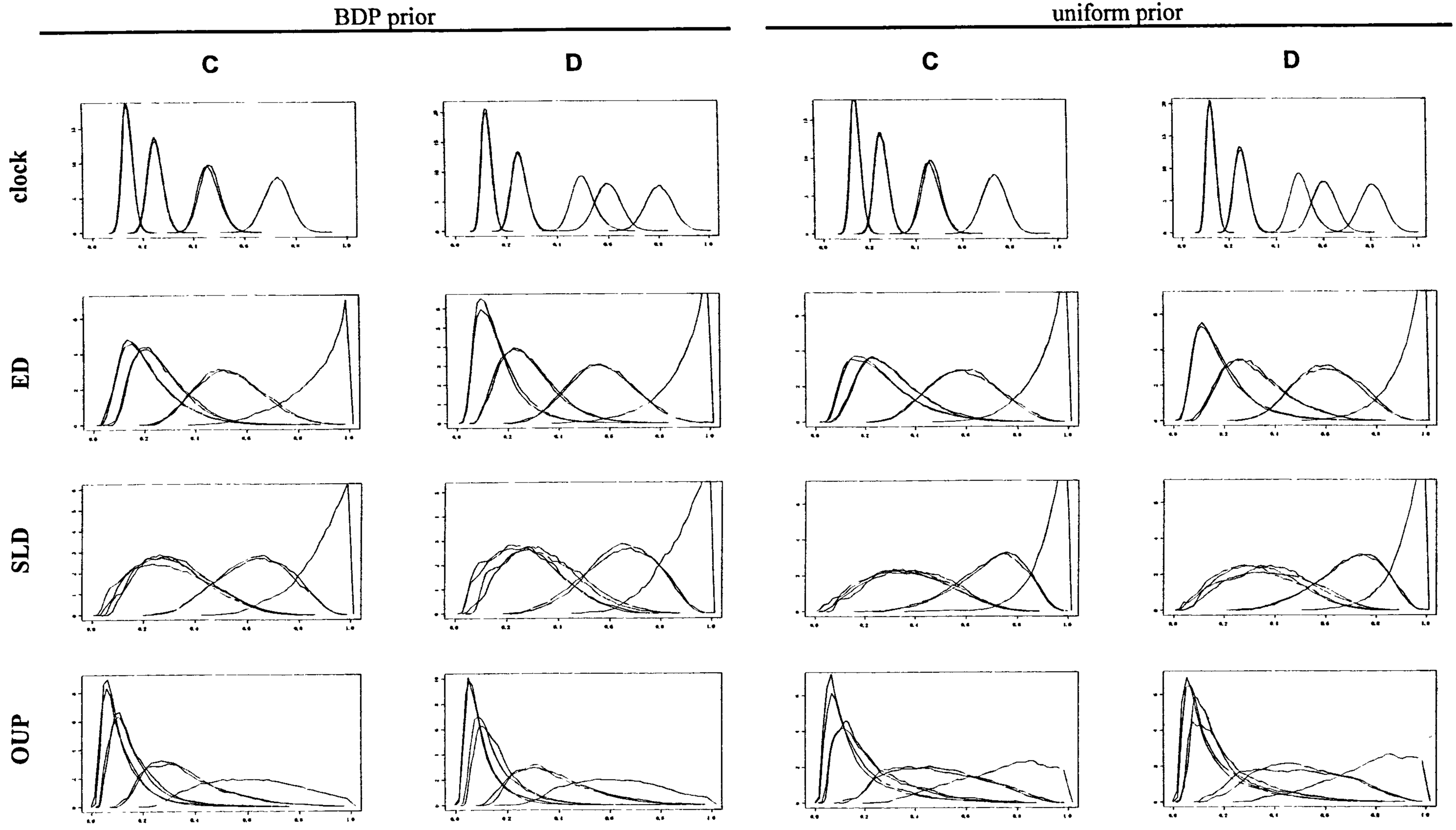


Figure II.14. Posterior distributions of the divergence times under the more complicated simulation trees. The different models of rate change are: the clock (first line), the ED model of rate change, SLD and OUP (last line).

model that gives the most imprecise rate estimates. This implies that the variance for the estimated divergence times may be underestimated under this model. While SLD gives good estimates under the clock, this model appears more sensitive to departure from purely autocorrelated rate changes than ED. Taken together, these results suggest that although ED is a very crude model, and that it does not yield the best fit on real data analyses (see section II.2 and Chapter III), it appears to be robust to violations of its underlying assumptions. This also seems to be the case for the BDP prior distribution on the model of speciation. Note that the time of the root is rarely correctly estimated. Considering the introduction of time constraints on a node – supposing there exist a well fossil-dated note – may improve the final estimates. This improvement has two aspects: (i) convergence would be better when the parameter is confined to a reduced parameter space and (ii) constrained intervals lead to better estimates than without constraints.

These simulations are still very simple. Some simplistic scenarios of rate change have been generated, under a simplistic substitution model. Real data analysis may provide us with a more practical test of the general model of rate change. This is what the next chapter is devoted to.

Chapter III

Bayes inference of times and rates: the Cambrian explosion revisited

Multicellular animals appear in the fossil record during a relatively short time interval, the "Cambrian explosion", around 560–530 million years ago (MYA) (Valentine et al., 1996). While the divergence date is necessarily older, molecular estimates are consistently older than 700 MYA (Wray et al., 1996; Feng et al., 1997; Gu, 1998; Wang et al., 1999; Bromham et al., 1998), hereby suggesting that more than half of the fossil record is missing (Benton, 1999) and that early Proterozoic evolution of multicellular animals was diffuse and undetected (Cooper and Fortey, 1998). However, these studies are based on the molecular clock hypothesis (Zuckerkandl and Pauling, 1965), which may not be the most appropriate and the most powerful when rates vary extensively. Here, a Bayesian approach has been implemented (Yang and Rannala, 1997; Thorne et al., 1998) to extend the results of Chapter II and to show that taking rate change and taxa sampling into account lead to estimates consistent with the fossil records. In particular, by analysing twenty-two genes, it is shown that the divergence of protostomes–deuterostomes (PD) is estimated to be around 587 ± 158 MYA. Moreover, most of the rate variation of the sampled genes appears to be distributed around two bursts of evolution, one around the end of the Precambrian, around the PD split, and one in the Silurian, corresponding to the differentiation of the vertebrates.

III.1 – Origin of the Metazoa: from the Cambrian explosion to the slow fuse hypothesis

III.1.a – The nature of the Cambrian explosion

The existence of a "Cambrian explosion" of metazoan fossils has long been known as it is already noted by Darwin (1859) – although this abrupt appearance was in conflict with his theory of *gradual* change. Numerous phyla are present in the early Cambrian (Benton, 1993), whereas mostly the enigmatic Ediacaran fauna is found before the base of the Cambrian, dated at 544 MYA (Bowring et al., 1993). Its exceptional preservation is

explained by the absence of predators and scavengers (Glaessner, 1984). This fauna was interpreted originally as a range of soft-bodied animals belonging to Cnidaria, Annelida and Athropoda; only one taxon had an unknown status (Glaessner, 1984). This was later challenged by Seilacher and co-workers (Seilacher, 1985; Buss and Seilacher, 1994) who suggested that organisms with a "quilted" body plan actually belong to an extinct high-level taxon, the Vendobionta. More recently, a study suggested that many species with this quilted body plan are more similar to each other than they are to any modern groups. This supports the concept of Vendobionta as a late Neoproterozoic group, consisting of multifoliate organisms with a distinctive quilted segmentation (Narbonne et al., 1997). Prior to this Ediacaran fauna, the only evidence of animals would be the presence of horizontal burrows, but these traces are often reinterpreted as pseudo-fossils (McMenamin, 1989).

What triggered this burst of morphological innovation is still debated, but recent studies have shed some new light. During the Cambrian, most of the environmental conditions, such as atmospheric composition, global temperature or geology, were similar to those prevailing today (Wallace, 1997, pp.77–78). The main difference was probably the absence of large animals and of sophisticated predators and / or parasites. The general scenario, at least for the "morphological Cambrian explosion", is then one of a proliferation into an empty niche (see Wallace, 1997, p.80), proliferation which occurred about 544 MYA, within a period of time as short as 10 to 15 MY (Bowring et al., 1993). In contrast to this situation during the Cambrian, the Precambrian differentiation of the Metazoa appears to have occurred in a context of geological and climatic upheavals. Recent paleomagnetic data indicated that continents moved at rates much higher than those observed today (Kirschvink et al., 1997), which led to the formation and destruction of two supercontinent (Hoffman, 1991). These fast Wilson cycles dramatically affected

the climate, with a series of global glaciations that occurred ca. 750–580 MYA, each lasting for millions of years and ending in scorching heat caused by an enrichment of atmospheric greenhouse gases (Hoffman, 1991; Hoffman et al., 1998). According to this "snowball Earth" hypothesis, conditions during the ice ages were such that no other life than bacteria is considered to have survived (Runnegar, 2000), except possibly under the conditions modelled by Hyde et al. (2000), which allow for an equatorial belt of open water. Disappearance of snowball conditions is thought have permitted the Cambrian morphological explosion (Kirschvink et al., 1997), probably in conjunction with other factors such as high oxygen levels that could make complex biochemical structures such as collagen possible (Knoll and Carroll, 1999).

The picture emerging, mainly from palaeontology but also from developmental biology, is that of multicellular animals originating no later than 613 MYA (Valentine, 1994; Valentine et al., 1996, 1999), first as the Vendobionta, which either were supplanted in the late Vendian by modern metazoans (Fedonkin, 1985), or underwent a dramatic decline in the late Vendian, with some lineages persisting until at least 510 MYA (Crimes et al., 1995). Although there is no consensus among palaeontologists, those favouring the hypothesis of a Phanerozoic metazoan life seem to locate the origin of the animals in the Vendian (Fortey et al., 1996), that is above the 613 MYA "lower bound". These authors argued that the evolution of the metazoan crow group taxa should have demanded an extensive period of time, during which metazoan life would have persisted as small animals, unlikely to fossilise. The Cambrian explosion may only reflect a sudden and simultaneous size increase in most of the groups, but the date of the origin is still contentious. Nonetheless, some recent isotopic dating suggests that the early Cambrian contains considerable time for their assembly and diversification (Knoll and Carroll, 1999).

This view is challenged by almost all molecular studies. Although the estimated divergence dates are very disparate, all these studies agree that the basal divergence between the protostomes and the deuterostomes occurred before 700 MYA, in the Proterozoic. Brown et al. (1972) provided one of the first estimate, based on the cytochrome c, at ~750 MYA; Runnegar (1982) then proposed a date around 850 MYA based on hemoglobin, re-evaluated to 700 MYA by using three genes (Runnegar, 1986); more recently the 18S and partial mitochondrial DNA sequences analysed by Bromham et al. (1998) suggested estimates as ancient as 2,250 MYA. This early-origin hypothesis seems all the more robust as it is supported by the analysis of a large number of genes: Wray et al. (1996) analysed five mitochondrial genes and three nuclear genes which gave dates ~1,000 MYA; Wang et al. (1999) considered a total of 75 genes to estimate different divergences, and dated this same protostomes / deuterostomes split to ~990 MYA with 50 genes. Similar results can be found in Feng et al. (1997) or Gu (1998). This deep-origin theory may have found some support in the recent find of some evidence of complex, multicellular animals found in sediments thought to have deposited 1,100 MYA (Seilacher et al., 1998). However, the age of the sediments has seriously been criticised (Brasier, 1998) as fossils of a typical early Cambrian fauna have been found in the immediately overlying sediments.

III.1.b – Late-arrival: the phylogenetic fuse hypothesis and its consequences

The main consequence of such a late-arrival hypothesis, with a genetic origin at ~1,000 MYA, is a decoupling of the evolutionary processes of (i) speciation and (ii) adaptation (Vermeij, 1996; Fortey et al., 1996). This separation in time of the two processes conducted some authors to formulate the phylogenetic fuse hypothesis (Cooper and Fortey, 1998; Easteal, 1999) – an extended period of evolutionary innovations that has left little or no fossil record. The recent discovery of a crustacean in early Cambrian

strata (511 MYA) from England (Siveter et al., 2001) provides some support to a phylogenetic fuse (Fortey, 2001), as such a differentiation is assumed to require a previous history of the group – note the slight circularity of the argument which tends to disregard the possibility of rapid innovations. This can be viewed as a rejoinder to Darwinian gradualism, in response to the punctuated equilibria theory proposed by Eldredge and Gould (1972).

In the light of these data, three hypotheses can be posited to explain the nature of the Cambrian explosion:

1 – lineage divergence (Vendian or pre-Vendian) occurred before evolution of multicellularity (Cambrian);

2 – lineage divergence (Vendian or pre-Vendian) occurred after evolution of multicellularity, but before evolution of a body size large enough to be fossilised and / or evolution of skeletal structures (Cambrian);

3 – lineage divergence, evolution of multicellularity, evolution of large body size and evolution of skeletal structures occurred simultaneously (Vendian or pre-Vendian), but fossilisation conditions were poor and no trace was left in the fossil record.

The similarity of the developmental stages characterising the animals until the moment they start to differ ontologically (the so-called phyletic stage) excludes the first hypothesis (Slack et al., 1993). Moreover, the Edicaran fauna makes the third hypothesis unlikely. Consequently, the results obtained so far in estimating the timeframe of the evolution of the Metazoa suggest that the Cambrian explosion is actually a rapid augmentation of body size paralleled by a morphological diversification. It must be noted that this rapid radiation occurred *simultaneously* in *all* the phyla. A coevolutionary arms race is usually proposed to explain this (Conway-Morris, 1998b).

In the next two sections of this chapter, I investigate a fourth hypothesis, according to which early lineage divergence (Vendian or pre-Vendian) might be a spurious result emerging from poor modelling.

III.2 – Dating the molecular origin of the Metazoa in a Bayesian framework

To test this new hypothesis, twenty-two genes were retrieved from the National Center for Biotechnology Information database (Genebank). They consist of eleven nuclear genes (18S rRNA; actin; α-tubulin; β-tubulin; calreticulin; catalase; elongation factor 1 [ef-1]; histone H1; heat shock protein 70 [hsp-70]; protein kinase C [pkc]; troponin C) and eleven mitochondrial genes (cytochrome c oxidase [cox] subunit I, II and III; cytochrome B [cytB]; NADH dehydrogenase [nad] subunit 1 to 6 and 4L). The corresponding accession numbers are given in Annex 2. These genes were chosen for (i) their extensive representation across the metazoa, (ii) including both protostome-deuterostome and echinoderm-chordate splits and (iii) for which at least one fossil calibration point was available for an age greater than 300 MYA in order to have points as close as possible to the Cambrian. Alignments were performed with ClustalW 1.8 (Thompson et al., 1994) with default settings and were checked by eye. The calibration points used are chosen from the fossil records (Bromham et al., 1998) (in MYA): Collembola–Pterygota, 390; Aranaea–Scorpionida, 405; Coelacanth–Dipnoi/Tetrapoda, 418; Osteichthyes–Dipnoi/Tetrapoda, 428; Asteroidea–Echinoidea, 500; Agnata–Gnathostoma, 510; Arachnida–Merostomata, 520; Cephalochordata–Chordata, 530. Each gene has between one and eight calibration points.

The Bayesian framework used is the one described in Chapter II. The only modification is the following. The hyperparameters of the process modelling rate change are integrated out assuming vague hyperprior distributions: lognormal of mean log(.5)

and variance .75 for the drift coefficient and gamma distribution of mean 15 and variance 25 for the diffusion coefficient. The hierarchical model can be summarised as:

$$
\begin{array}{lll}
\text{data level} & X \sim \ell(\theta \mid X) & \text{– likelihood of the parameters } \theta \\
\text{parameters' level} & \theta \sim p(R, T \mid \lambda, \mu, \rho, \beta, \sigma^2) & \text{– models for rates } R \text{ and times } T \\
\text{prior level} & \lambda \sim \mathrm{U}(0, 15) & \text{– hyperpriors of model for } T \\
 & \mu \sim \mathrm{U}(0, 15) & \\
 & \rho \sim \mathrm{U}(0, 0.001) & \\
 & \beta \sim \text{lognormal}(\log(0.5), 0.75) & \text{– hyperpriors of model for } R \\
 & \sigma^2 \sim \text{gamma}(9, 0.6) &
\end{array}
$$

For each gene, the marginal posterior distributions $p(T \mid X)$ and $p(R \mid X)$ are approximated by the means of a Markov chain Monte Carlo algorithm based on the Metropolis–Hastings sampler (Gilks et al., 1996). As in Chapter II, at each step of the Markov chain, a new state θ^* is proposed for parameters θ of the model (divergence time, rate, κ or α). This state is accepted with probability $\min\{1, p(\theta^* \mid X) / p(\theta \mid X)\}$. The 50,000 first steps or the chain are discarded (burn-in) and each chain is then sampled every 500 steps until 10,000 samples are collected. Convergence is checked by running four short preliminary runs for each gene under each model, analysing time series outputs for each parameter and checking consistency of the estimates across the different runs. Inferences are based on the median of each marginalized parameter. Model selection is based on the posterior Bayes factor (*PBF*) of the hierarchical model defined above.

III.3 – Bayesian large-scale analysis under models of rate change

I propose here to relax the molecular clock assumption in a Bayesian framework to estimate the mode and tempo of molecular evolution of the Metazoa. The eleven nuclear genes and eleven mitochondrial genes analysed have an average number of 26 taxa and 1,388 nucleotides per gene (Table III.1). For each gene, the tree topology used is the species tree, assumed to be known (Nielsen, 1995). For the purpose of dating, each tree is rooted by either a land plant (Arabidopsis), a fern (Polypodium for the 18S rRNA gene) or a fungi (Schizosaccharomyces for the troponin *c* gene). In order to reflect the most basal split (Parazoa–Eumetazoa), a diploblastic animal (Cnidaria) is included in the analysis whenever it is possible. To reduce errors associated with calibration points, only fossil-based dates were considered, and as many calibration points as possible were used (up to eight for the 18S rRNA genes). I focus on two key transitions (Conway-Morris, 1998a): the protostome–deuterostome (PD) divergence, which marks the appearance of "higher Metazoa" (Eumetazoa), and the echinoderm–chordate (EC) divergence, as it predates the origin of the vertebrates.

The molecular clock hypothesis is tested using two approaches: the likelihood ratio test (LRT) and the posterior Bayes factor (see Chapters I and II). For all the genes studied, the molecular clock hypothesis was strongly rejected (Table III.1–2). Date estimates of the PD divergence under the clock are above 700 MYA for most genes (Table III.2), with an average (±2 SE) of 1,113 ± 832 MYA, i.e. before the Vendian (Figure III.1A). These results are in agreement with previous molecular studies (see above). Different genes produced substantially different estimates for the PD divergence, ranging from ~500 MYA for calreticulin to ~1,990 MYA for Cox1 (Figure III.1A).

To relax the molecular clock assumption, we implemented two models of rate change over time: the exponential model and the Ornstein-Uhlenbeck process (Chapter II). Table

Table III.1. Genes sampled for analysing the timing of the origin of the Metazoa with the likelihood ratio test statistic of the molecular clock hypothesis.

Location	Gene	NS	SL	NCP	$2\,\delta_{LRT}$
mitochondrial	cox1	35	1681	4	2341.22
	cox2	35	804	4	740.33
	cox3	35	813	4	1173.48
	cytB	35	1285	4	1847.06
	nad1	35	1087	4	1336.89
	nad2	35	2589	4	747.74
	nad3	35	403	4	312.29
	nad4	35	1563	4	1110.97
	nad4L	35	319	4	233.19
	nad5	34	2073	4	964.33
	nad6	35	655	4	348.79
nuclear	18S	40	1032	8	867.50
	actin	14	1135	1	73.36
	α tubulin	21	1365	1	460.26
	β tubulin	18	1389	1	219.65
	calreticulin	12	1424	1	111.17
	catalase	11	2379	1	776.12
	ef-1	30	2562	2	4840.38
	histone H1	13	743	1	380.29
	hsp-70	12	2026	1	163.99
	pkc	11	2633	1	724.87
	troponin C	13	573	1	163.38

Notes –NS: number of sequences. SL: sequence length. NCP: number of calibration points. $2\,\delta_{LRT}$: likelihood ratio test statistic (minus twice the likelihood score difference).

III.2 shows that the two models give similar estimates, and the PD divergence is dated at 587 ± 158 MYA and 580 ± 112 MYA, respectively. The effect of relaxing the molecular clock is seen to be dramatic, and estimates from the two models of rate change (Figure III.1, B and C) are found to be consistent with paleontological data (Valentine et al., 1999).

Before drawing a firm conclusion, I examine the effects of several factors. First, the exponential model used is very crude and unrealistic, while the Ornstein-Uhlenbeck process is more complex. Table III.2 shows that the latter model fitted the data better than the former for most genes (median log PBF_{OE} = 16). Both models clearly outperformed the clock model (e.g., log PBF_{OC} >> 10, Table III.2) and gave consistent estimates of divergence dates (Table III.2, Fig. III.1, B and C). Therefore, the above estimates of the PD and EC splits are robust to the specification of the model of rate change, and may not be improved by more complex models. Second, when the Markov chain is run with no data, the node corresponding to the PD split has median prior divergence times ~500 MYA. This is the case under any prior distribution for rates, including the clock. As posterior estimates under the clock are almost twice as large, the priors chosen are, if not absolutely uninformative, vague enough to allow a correct estimation. Third, it is becoming widely accepted that uncertainties with respect to some parameters of a model of evolution affect the estimates (Huelsenbeck et al., 2000b). For instance, overlooking among-site rate variation biases divergence date estimates. I have addressed this issue by integrating the hyperparameters out of the sub-models for nucleotide substitution. Fourth, one important parameter of the prior model describing the speciation process is the sampling fraction. The analysis (Table III.2, Fig. III.1) assumes that the sampling fraction has a uniform prior distribution U(0, ρ_{up}) with the upper bound ρ_{up} = .001. Larger ρ_{up}

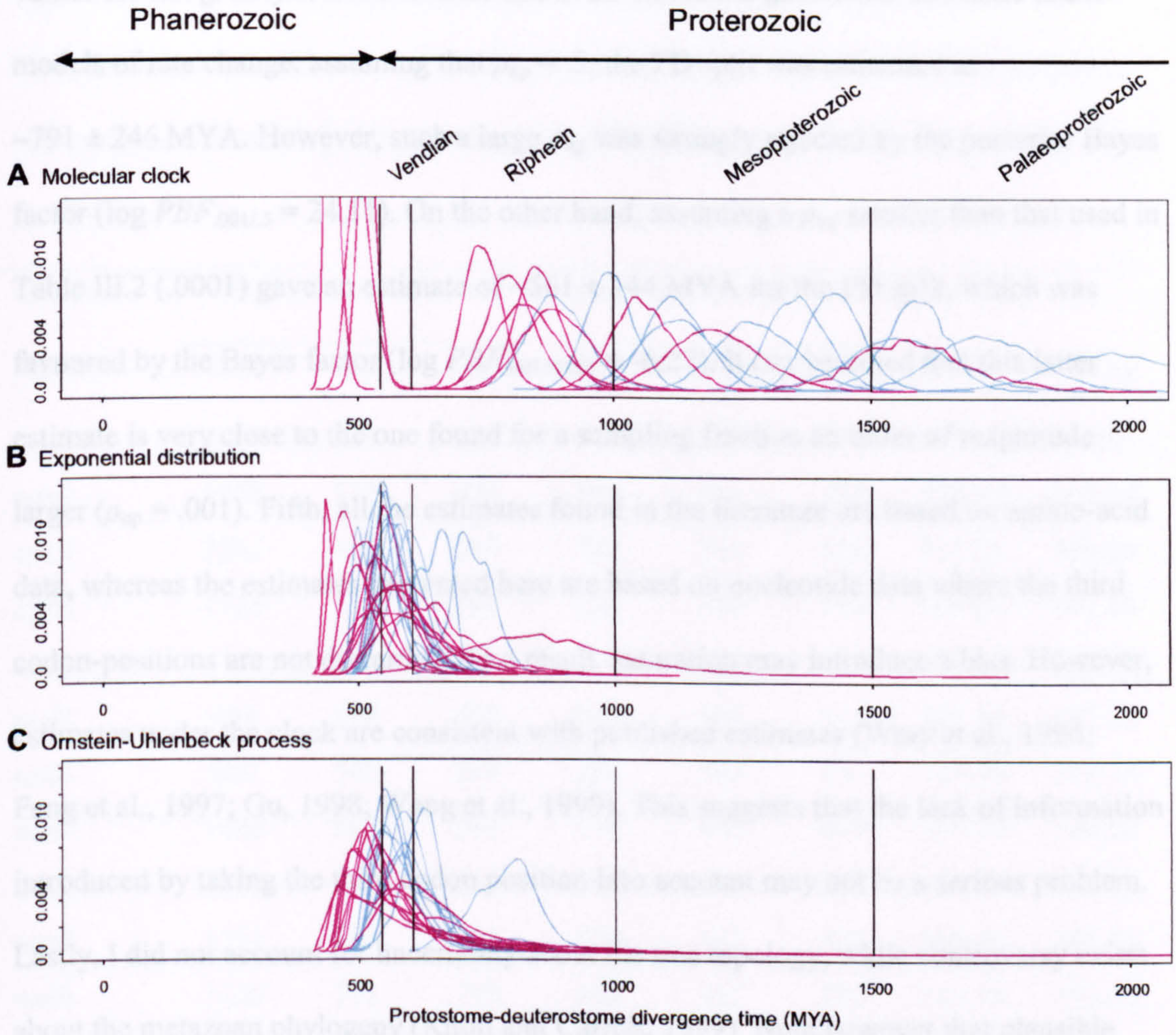


Figure III.1. Posterior distributions of the divergence time between protostomes and deuterostomes. Eleven nuclear genes (magenta) and eleven mitochondrial genes (blue) were analyzed under three models of rate change: **A** the Bayesian molecular clock, **B** the exponential distribution, **C** the Ornstein-Uhlenbeck process.

values did not greatly affect estimates under the clock, but gave older estimates under models of rate change: assuming that $\rho_{up} = .5$, the PD split was estimated at ~791 ± 246 MYA. However, such a large ρ_{up} was strongly rejected by the posterior Bayes factor (log $PBF_{.001/.5} = 24.11$). On the other hand, assuming a ρ_{up} smaller than that used in Table III.2 (.0001) gave an estimate of ~561 ± 144 MYA for the PD split, which was favoured by the Bayes factor (log $PBF_{.001/.0001} = -6.27$). It can be noted that this latter estimate is very close to the one found for a sampling fraction an order of magnitude larger ($\rho_{up} = .001$). Fifth, all the estimates found in the literature are based on amino-acid data, whereas the estimates presented here are based on nucleotide data where the third codon-positions are not discarded. As a result, saturation may introduce a bias. However, estimates under the clock are consistent with published estimates (Wray et al., 1996; Feng et al., 1997; Gu, 1998; Wang et al., 1999). This suggests that the lack of information introduced by taking the third codon position into account may not be a serious problem. Lastly, I did not account for uncertainty about the tree topology, while controversy exists about the metazoan phylogeny (Knoll and Carroll, 1999). Note however that plausible topologies gave similar speciation date estimates (Yoder and Yang, 2000).

It is of interest to examine whether the Cambrian explosion, as recorded by the fossils, has been preceded by a burst of molecular evolution (Bromham and Hendy, 2000). Fig. III.2 summarizes the estimates of relative rates against time from the exponential model of rate change. I define elevated relative rates as those greater than the 95th percentile of the distribution of relative rates over branches and over the sampled genes (rightmost panel of Fig. III.2). High relative rates occur mainly between ~640 MYA (late Riphean) and ~420 MYA (Silurian). The average relative rate is almost twice as large during this period (1.37) than either before (0.71) or after (0.62) it.

Table III.2. Divergence times of two major clades (MYA). Speciation is modelled by a birth-death process with average species sampling of 0.05%. Dates are averaged over independent calibration points based on fossil data only.

	proto–deuterostome			echinoderm–chordate				
Gene	clock	exp	oup	clock	exp	oup	PBF_{OC}	PBF_{OE}
cox1	1988	538	624	810	506	544	1070.03	50.89
cox2	1278	577	602	799	513	533	324.29	165.14
cox3	1561	540	595	782	514	539	515.31	479.28
cytB	1724	572	583	707	508	512	785.07	807.64
nad1	1358	552	574	1329	504	538	579.24	380.82
nad2	994	562	558	949	500	511	333.19	482.39
nad3	889	569	599	634	500	513	160.98	319.14
nad4	1428	605	583	736	507	499	457.90	386.00
nad4L	1128	588	562	757	511	517	105.86	626.12
nad5	1591	650	603	829	513	532	386.46	467.11
nad6	1120	721	790	660	517	563	164.96	795.13
18S	1576	560	567	1318	549	546	271.60	-2.19
actin	888	492	533	593	457	475	30.87	134.74
α tubulin	799	516	528	683	505	492	196.91	-3.00
β tubulin	1197	579	583	706	522	473	102.68	-2.30
calreticulin	504	482	533	480	470	480	52.09	141.37
catalase	1057	579	531	1053	566	491	72.44	218.76
ef-1	854	602	527	850	594	524	753.71	714.50
histone H1	450	554	546	443	541	515	165.11	154.28
hsp-70	744	582	574	735	567	528	-80.62	114.65
pkc	519	638	539	514	628	518	198.33	255.75
troponin C	831	856	624	542	454	442	167.66	694.41

Notes –Posterior Bayes factors *PBF* are given on a $\log_{10}$-scale. PBF_{OC} is the ratio $PBF_{OUP,CLOCK}$; PBF_{OE} is the ratio $PBF_{OUP,EXP}$.

Elevated rates are mainly for branches prior to the PD, the EC and the Agnatha–Gnathostoma (jawless and jawed vertebrates) divergences. The last two are contiguous and may belong to a single long period of elevated rates. Because of our limited sampling around the Parazoa–Eumetazoa split, it is difficult to detect any such burst at that time. It is remarkable that these bursts of evolution have concerned most of the genes, suggesting genome-wide phenomena, which could correspond to major duplication events (Pollard and Holland, 2000; Miyata and Suga, 2001; Abi-Rached et al., 2002). Other lineages with high rates (Fig. III.2) belong to the invertebrates, but here high rates are more gene-dependent. Subsequent "bursts" of evolution (< 400 MYA) are smaller in magnitude and mainly concern, at least in our restricted species sampling, parasites.

The results presented in this chapter show that the evolutionary history of the Metazoa has been complex, with at least two major bursts of molecular evolution. Furthermore, differentiation has been explosive, in the sense that our date estimates suggest it occurred in a relatively short period of time, as the PD and EC splits were found to be separated by an average of 67 MY. I also found this period of time to be characterised by elevated molecular evolutionary rates in most of the genes analysed. Previous studies suggested possible genome-wide duplication events around these divergences (Pollard and Holland, 2000; Miyata and Suga, 2001; Abi-Rached et al., 2002), which might have lead to relaxed selective constraints and high molecular evolutionary rates. The environmental elements that could have triggered the Cambrian explosion remain unclear (Knoll and Carroll, 1999), but dates estimated under models of rate change indicate a probable origination of the Metazoa at about the Varanger ice age (~620–580 MYA), hereby renewing interest into possible refugia during a snowball Earth (Hyde et al., 2000). It is important to realise that the Cambrian explosion does not

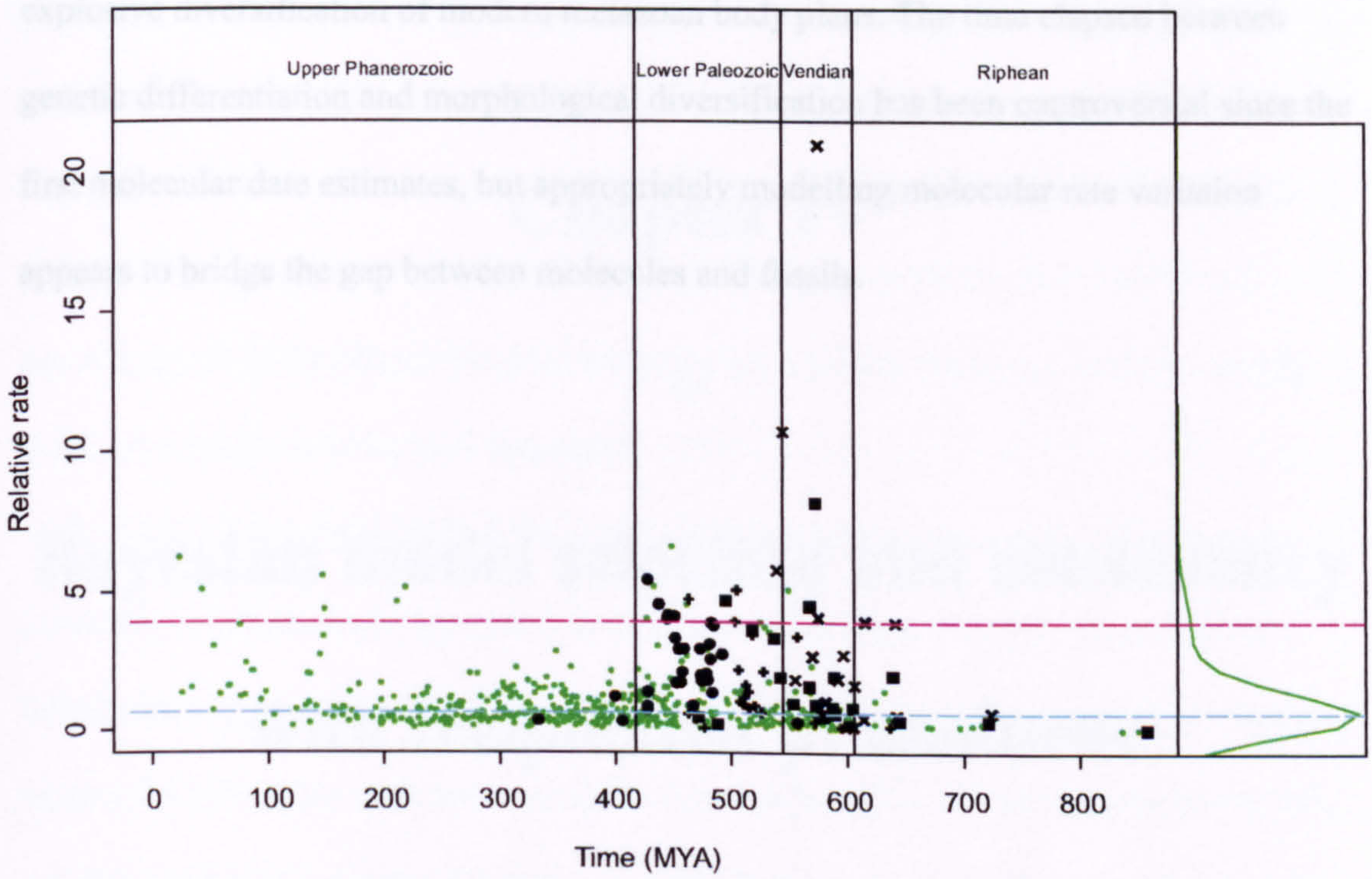


Figure III.2. Relative rates of evolution plotted against the estimated divergence dates for the twenty-two genes studied. For each gene the relative rate is calculated by the estimated rate for the branch divided by the average rate for the gene. The x-axis indicates the age of the descendent node of the branch. Horizontal lines correspond to the median (blue) and to the 95th percentile (magenta) of the distribution of relative rates over branches and genes. Four branches leading to the following divergences are indicated as: Parazoa–Eumetazoa split (×); PD split (■); EC split (+); split basal to vertebrates (●). A schematic geological scale is also given (topmost).

represent the origin of the animals (Knoll and Carroll, 1999), but instead marks the explosive diversification of modern metazoan body plans. The time elapsed between genetic differentiation and morphological diversification has been controversial since the first molecular date estimates, but appropriately modelling molecular rate variation appears to bridge the gap between molecules and fossils.

Chapter IV

Bayesian model selection and consistency with frequentist procedures

Random and systematic errors make reconstruction of phylogenetic trees uncertain, in the sense that the estimated tree may not be the true tree. Recently, two bootstrap-based tests were implemented to overcome some statistical issues: the non-parametric Shimodaira-Hasegawa (SH) test and the parametric Swofford-Olsen-Waddell-Hillis (SOWH) test. However, their application lead to strikingly different conclusions, the SOWH test appearing extremely liberal. To date, it is not clear whether this is due to the technique used, or to the form of the null hypothesis.

In order to understand this difference, I present here four new tests. Based on two contrasted approaches (non-parametric bootstraps and Bayes factors), these tests are formulated either in the framework of hypothesis tests or as significance tests. I show that frequentist and Bayes approaches can give consistent results within a testing framework, and that the SOWH and the SH tests can respectively be interpreted as hypothesis and significance tests. Hence, the observed difference between the two tests, SH and SOWH, does not amount to the approach or technique used.

IV.0 – Hypothesis tests vs. significance tests of trees

A phylogenetic tree represents an evolutionary hypothesis and consists of two components: the amount of evolution separating the taxa under study, that is the branch lengths, and the branching pattern, or topology. When estimating a tree, it is important to know how reliable our estimate is. In the framework of maximum likelihood (Felsenstein, 1981), statistical tests of competing trees have been available for a decade (see Hillis et al., 1996). Kishino and Hasegawa (1989) and Hasegawa and Kishino (1989) first proposed a test to evaluate whether two randomly chosen topologies differ significantly. Recent implementations of this test, traditionally denoted KH, approximate the confidence interval (CI) of the log-likelihood difference by assuming it is normally

distributed under the Central Limit Theorem. As noted by Shimodaira and Hasegawa (1999), this approach is similar to one found in the statistical (Linhart, 1988) or the econometrical (Vuong, 1989) literature. Under the null hypothesis, the two trees are not significantly different and the expected log-likelihood difference is zero. If the observed difference is greater than a critical value, represented by the CI of size $1 - \alpha$, then the two models are declared significantly different at level α.

Recently, three general warnings were issued about usage of the KH test (Shimodaira and Hasegawa, 1999; Goldman et al., 2000). First, tree topologies are commonly tested against the maximum likelihood (ML), which is a posteriori selected. By definition, the ML tree has the highest likelihood. Therefore, the null hypothesis is not that the expected log-likelihood difference between the ML tree and the other one is zero, $E(\Delta\ell) = 0$, but rather that $E(\Delta\ell) \leq 0$ (see section IV.1 for a more exact definition). Second, the practice just described is not only carried out on one tree, but is often performed to evaluate the ML tree against a collection of alternative trees. Consequently, multiple comparisons are performed. These two issues may lead to rejecting the null hypothesis too often, making the KH test liberal. Third, the CI's of the KH test were originally described for a two-tailed test, when they are often used in the context of an upper-tailed test. The effect of this point depends on the *p*-value obtained. However, in most cases the KH test is found to be too liberal (Shimodaira and Hasegawa, 1999; Shimodaira, 2001).

Shimodaira and Hasegawa (1999) were the first to highlight the difficulties associated with the KH test. These authors proposed a non-parametric test, the SH test. Regarded as a modified KH test, the SH test assesses which model (substitution process and tree T_i) is better and makes proper allowance for multiple comparisons and a posteriori selected trees (Shimodaira, 1998). It is acknowledged that the test is conservative (Shimodaira and Hasegawa, 1999), but it is not clear to what extent.

More recently, Goldman et al. (2000) implemented a parametric bootstrap, the SOWH test, described by Swofford et al. (1996) and named after these authors. The test evaluates a tree T_i against an alternative T_j, usually a posteriori selected. The null hypothesis is that T_i is the correct tree (H_i^0: T_i) that is, the tree that would be estimated with infinite sequence length. The alternative T_j is here explicitly specified, and is in practice taken at the ML tree, T_{ML}. The idea underlying the parametric bootstrap is to estimate the distribution of the log-likelihood difference $\Delta\ell$ between T_{ML} and T_i. The *p*-value is then the probability that difference $\Delta\ell$ generated under H_i^0 is larger than the observed $\Delta\ell$. To this effect, sequences are simulated under H_i^0, and the likelihood of each replicate is optimised over the parameter space and the tree space. This test was also described by Huelsenbeck and Crandall (1997), who previously implemented a related algorithm (Huelsenbeck et al., 1996) given by Waterman (1995: pp. 374–375), where the parametric replicates are simulated under T_{ML}, not T_i. However, the null hypothesis is H^0: T_{ML} in this case, which is not conventional since the null hypothesis is typically the one we try to reject. In practice, the SOWH test is found to result in *p*-values generally smaller than under the SH test (Goldman et al., 2000). Three possible causes for such a discrepancy were proposed: (i) different form of the null hypothesis, (ii) greater power of parametric tests and (iii) their reliance on a substitution model (Goldman et al., 2000, p. 669).

To assess these three suggestions, I have implemented four new tests of phylogenetic trees within two contrasted approaches: frequentist and Bayesian. In each approach, I make the distinction between two frameworks: (i) significance tests quantify the evidence against the tested trees, whereas (ii) hypothesis tests measure in a pairwise manner the evidence against a specific tree in presence of an alternative (Table IV.0). These tests differ by their respective null hypotheses which are, respectively, that (i) all the tested

Table IV.0. Summary of the tests (T) implemented in this chapter, with their respective mnemonics and statistics: significance (S) or hypothesis (H) tests are formulated either in a frequentist (F) or a Bayes (B) approach. See text for the exact definitions.

Framework	Frequentist approach	Bayes approach
(i) – significance tests	TFS: $\widetilde{p}_i^{\,\mathrm{lfd}}$	TBS: $BF_{\bar{T},i}$
(ii) – hypothesis tests	TFH: $\widetilde{p}_i^{\,\mathrm{ML}}$	TBH: $BF_{ML,i}$

trees are equivalent, or (ii) the ML tree is not the correct tree. They are also distinguished by the absence (i) or the presence (ii) of an explicitly stated alternative hypothesis. Both frameworks lead to the construction of confidence sets of trees by inverting the test procedure.

These two frameworks are naturally implemented in a frequentist approach, and I show how it can be done by means of a non-parametric bootstrap that adjusts *p*-values for multiple comparisons. In the Bayesian approach, I compare two measures: the posterior probability of a tree, which assesses the uncertainty regarding the tree (Yang and Rannala, 1997; Larget and Simon, 1999), and the Bayes factor (BF), which is the ratio of probabilities of the data under two models. The new frequentist and Bayesian tests are illustrated with two examples, one using DNA sequences from HIV-1 isolates (Goldman et al., 2000), and one using the complete mitochondrial DNA genome of six mammals. These examples suggest that the difference between the SH and SOWH tests is mainly due to the testing framework (significance vs. hypothesis test), and not to the approach (frequentist vs. Bayes) or to the technique (parametric vs. non-parametric) used.

IV.1 – The frequentist approach: p*-value adjustments by non-parametric bootstrap*

Comparing a tree T_i with the ML tree T_{ML} can be done by testing the particular null hypothesis H_i^0: $E_\theta\{\ell(T_{ML}) - \ell(T_i)\} \leq 0$, where $\ell(T_i)$ denotes the log-likelihood function at the ML estimate of all the parameters θ (branch lengths and parameters of the substitution model). When H_i^0 is rejected at the significance level α, T_{ML} is significantly better than T_i. When several trees are tested against T_{ML}, multiple tests are performed, and the probability of rejecting a particular null hypothesis is larger than the prespecified significance level α. This risk of false discovery, called the Familywise Error Rate (FWE), is corrected by making tests more conservative. More precisely, the FWE is

defined as the probability of rejecting at least one H_i^0 given that the complete null hypothesis $H_C^0 = \bigcap_{i=1}^{k} H_i^0$ is true, i.e. all the trees T_i are at least as good as T_{ML}:

$$\text{FWE} = \Pr(\text{Reject at least one } H_i^0 \mid H_C^0) \qquad \text{(IV.1)}$$

Generally, the FWE is said to be controlled if FWE $\leq \alpha$. A convenient way to achieve such a control is by *p*-value adjustments. This can be done using the Bonferroni correction: when performing k tests, H_i^0 is rejected when the *p*-value is less than α / k. It can be shown (e.g., Westfall and Young, 1993, p. 44) that if we assume that the *p*-values are random variables P_is uniformly distributed on U[0,1] under their respective null hypotheses, the Bonferroni correction controls the FWE at level α. Bar-Hen and Kishino (2000) recently used this correction in phylogenetics. However, for highly correlated data, as it is the case with molecular sequences, the FWE is no longer controlled. The Bonferroni correction is usually conservative for data with light-tailed sampling distributions, but can become liberal for heavy-tailed distributions (see Westfall and Young, 1993, p. 44). Resampling techniques take the distributional characteristics of the data into account, hereby producing smaller adjusted *p*-values and increasing the power of the correction. In practice, adjusted *p*-values, denoted $\tilde{p}_i$ hereafter, are defined as the smallest significance level for which H_i^0 is still rejected at FWE α, or:

$$\tilde{p}_i = \Pr(\min_{1 \leq j \leq k} P_j \leq p_i \mid H_0^C) \qquad \text{(IV.2)}$$

This procedure controls the FWE at level α, at least approximately (Westfall and Young, 1993, p. 53). Following this approach, I first present the construction of confidence sets of trees by the inversion of the significance testing procedure, before proceeding to hypothesis tests about the ML tree. Computational details of the test statistics and of the algorithm used are then given in a third subsection.

IV.1.a – Significance test of trees (TFS)

The complete null hypothesis H_C^0 defined above is actually a composite hypothesis, as the data X is tested against several distributions (log-likelihood functions ℓ_{Ti}'s for trees T_i's) where some parameters are left unspecified. The method generally employed to extend the theory of simple hypothesis testing to composite hypotheses is to reduce H_C^0 to a simple hypothesis. This is done by considering a weighted average of the distributions of H_i^0 over the tree space T i.e., $\int_T \ell(T_i)\, d\lambda(T)$ where λ is a probability distribution over T (Lehmann, 1959, p. 91; Kempthorne and Folks, 1971, p. 354). The choice of λ comes from the following consideration: the above average must reflect H_C^0 in that it must convey no information regarding the different trees. To this effect, an equal weight is assigned to the different distributions of H_C^0 by means of a uniform distribution λ, called the least favourable distribution (lfd) of H_C^0 at level α. In practice H_C^0 is reduced to the simple hypothesis, which is the average of the log-likelihoods $\ell(T_i)$ over the k tested trees: $E_k\{\ell(T_i)\}$. This critical value is denoted c_λ and the corresponding adjusted p-values obtained by the bootstrap procedure detailed below are denoted $\tilde{p}_i^{\,\text{lfd}}$. Note however that in practice $E_k\{\ell(T_i)\}$ is the average over the set of tested trees, whereas the distribution λ is defined over the tree space. Consequently, the power of this test is not uniform, as it depends on the trees tested. Yet, if the set of tested trees is large and includes the most likely trees, the influence of the remaining trees should be negligible (see the Bayesian argument below).

IV.1.b – Hypothesis test about the ML tree (TFH)

Significance tests do not consider an explicitly stated alternative hypothesis and thereby do not attach any value to the possibility that one of the tested trees might be the correct tree. The object of hypothesis tests is to select the correct tree, specifically assessing a given tree T_i against T_{ML}. This new objective is reflected by the choice of λ when

reducing the composite null hypothesis to a simple one: knowing that T_{ML} is the best tree, a peculiar weight is assigned to this tree. In the most extreme case, the critical value c_λ is set to $\ell(T_{ML})$. The corresponding adjusted p-values are denoted $\tilde{p}_i^{\,ML}$ hereafter. Note that unlike $\tilde{p}_i^{\,lfd}$, $\tilde{p}_i^{\,ML}$ is by construction not sensitive to the size of the set of tested trees.

IV.1.c – Computation

The test statistics t_i's are computed for the k alternative trees as $\{ \ell(T_i) - c_\lambda \} / s_i$. For each tree T_i, the standard error (SE) s_i is estimated as the SE of the distribution of sitewise log-likelihood values, as implemented in PAML (Yang, 1997b). The general form of this test statistic is close to the one of the normal test of the generalised likelihood ratio test (see Vuong, 1989, p. 318). More specifically, the test statistic used for significance tests (tests at the lfd) is:

$$t_{lfd} = \frac{\ell(T_i) - c_\lambda}{s_i} = \frac{\ell(T_i) - \sum_{j=1}^{k} \ell(T_j)/k}{s_i} \qquad \text{(IV.3a)}$$

while the statistic used for hypothesis testing of tree T_i against T_{ML} is:

$$t_{ht} = \frac{\ell(T_i) - c_\lambda}{s_i} = \frac{\ell(T_i) - \ell(T_{ML})}{s_i} \qquad \text{(IV.3b)}$$

This is "step 0" of the algorithm (Figure IV.1).

The general resampling algorithm used to estimate the distribution of the test statistics t_i's and adjust the p-values for multiple comparisons of k tree topologies follows Westfall and Young (1993, p. 47) and in schematised in Figure IV.1 **B**,**C**:

1. Initialise the counting variables: $C_i = 0$, $i = 1,\ldots, k$.
2. Generate a p-value vector $p^* = (p_1^*,\ldots, p_k^*)'$ from the same distribution as the original p-values under the complete null hypothesis. This step consists of three parts: (2-i) generate bootstrapped data X^* (e.g. 10,000 replicates), (2-ii) optimise branch lengths and parameters of the substitution model for each of the k topologies tested and (2-iii)

compute the p-values p^* from the simulated data as $\{ \ell(T_i^*) - \ell(T_i) \} / s_i^*$ where asterisks indicate ML estimates obtained from the bootstrapped data.

3. If $\min_{1 \leq j \leq k}(p_j^*) \leq p_i$, then increment C_i of one (i.e. C_i ++).
4. Repeat steps 2 to 3 N times. The estimated value of the adjusted p-value $\tilde{p}_i$ is approximated by C_i / N.

Note that the statistic $\{ \ell(T_i^*) - \ell(T_i) \}$ is used in step (2-ii) because we want to estimate the distribution of $\{ \ell(T_i) - c_\lambda \}$, the difference between our estimate of the log-likelihood and the unknown true tree (under H_0). Indeed, this distribution is estimated by the distribution of the difference between the resampled estimates and the original estimates. This corresponds to the first guideline by Hall and Wilson (1991), as applied e.g. by Efron et al. (1996) (see Chapter I.3.c). Furthermore, the test is based on the distribution of $\{ \ell(T_i^*) - \ell(T_i) \}/s_i$, which is the "bootstrap pivoting" of the second recommendation of Hall and Wilson (1991).

Care must be taken when resampling the original data X to respect the complete null hypothesis (the tested trees cannot be distinguished). In particular, in the presence of gene partitions, the resampling must be stratified (e.g. Yoder and Yang, 2000). Note that the set of the k-compared topologies is fixed, so that only real-valued parameters need to be updated. However, because of this optimisation step (2–ii), the algorithm described above would be very expensive in terms of required computation time. A time-saving approximation using the RELL approach (Kishino et al., 1990) is implemented in the following algorithm, where the sitewise log-likelihood values computed from the original data set X are bootstrapped rather than X itself:

1. Initialise the counting variables: $C_i = 0$, $i = 1,\ldots, k$.

2. Generate a p-value vector $p^* = (p_1^*,\ldots, p_k^*)$'. This step consists of two parts: (2-i) bootstrap the sitewise log-likelihood values (10,000 replicates) and (2-ii) compute the p-values p^* from the simulated data as $\{ \ell(T_i^*) - \ell(T_i) \} / s_i^*$ where asterisks indicate RELL estimates. When genes are combined, the RELL procedure is stratified within each partition of the data set.
3. If $\min_{1\le j\le k}(p_j^*) \le p_i$, then increment C_i of one (i.e. C_i ++).
4. Repeat steps 2 to 3 N times. The estimated value of the adjusted p-value $\tilde{p}_i$ is approximated by C_i / N.

The program `baseml` in PAML (Yang, 1997b) is used to compute the sitewise log-likelihood values. The second algorithm described above can be used to compute standard tests such as the SH and the SOWH tests. First, when $s_i = s_i^* = 1$, the $\tilde{p}_i^{\,\text{lfd}}$'s correspond to the p-values of the SH test; otherwise it is analogous to the weighted SH test (see remark 4 in Shimodaira and Hasegawa, 1999). Second, when only two trees are compared in the framework of hypothesis tests, such as T_1 against T_{ML}, steps 1–4 in Figure IV.1 reduce to:

$$X^* \rightarrow (t_1^*) \rightarrow \text{if}\{\, t_1^* \le t_1 \,\} \text{ then } \quad C_1 \text{ ++}$$

and the p-value is estimated as C_1 / N when drawing N replicates.

IV.2 – The Bayes approach I: Posterior probability of a tree

In molecular phylogenetics, evolutionary models have two components: the topology, and the substitution model for amino acid or nucleotide sequences. For instance, under the HKY85 + Γ nucleotide substitution model (Hasegawa et al., 1985; Yang, 1994b), the model not only includes the transition-transversion rate ratio κ, the among-site rate variation parameter α, and the base frequencies $\boldsymbol{\pi}$, but also the evolutionary tree T, and its

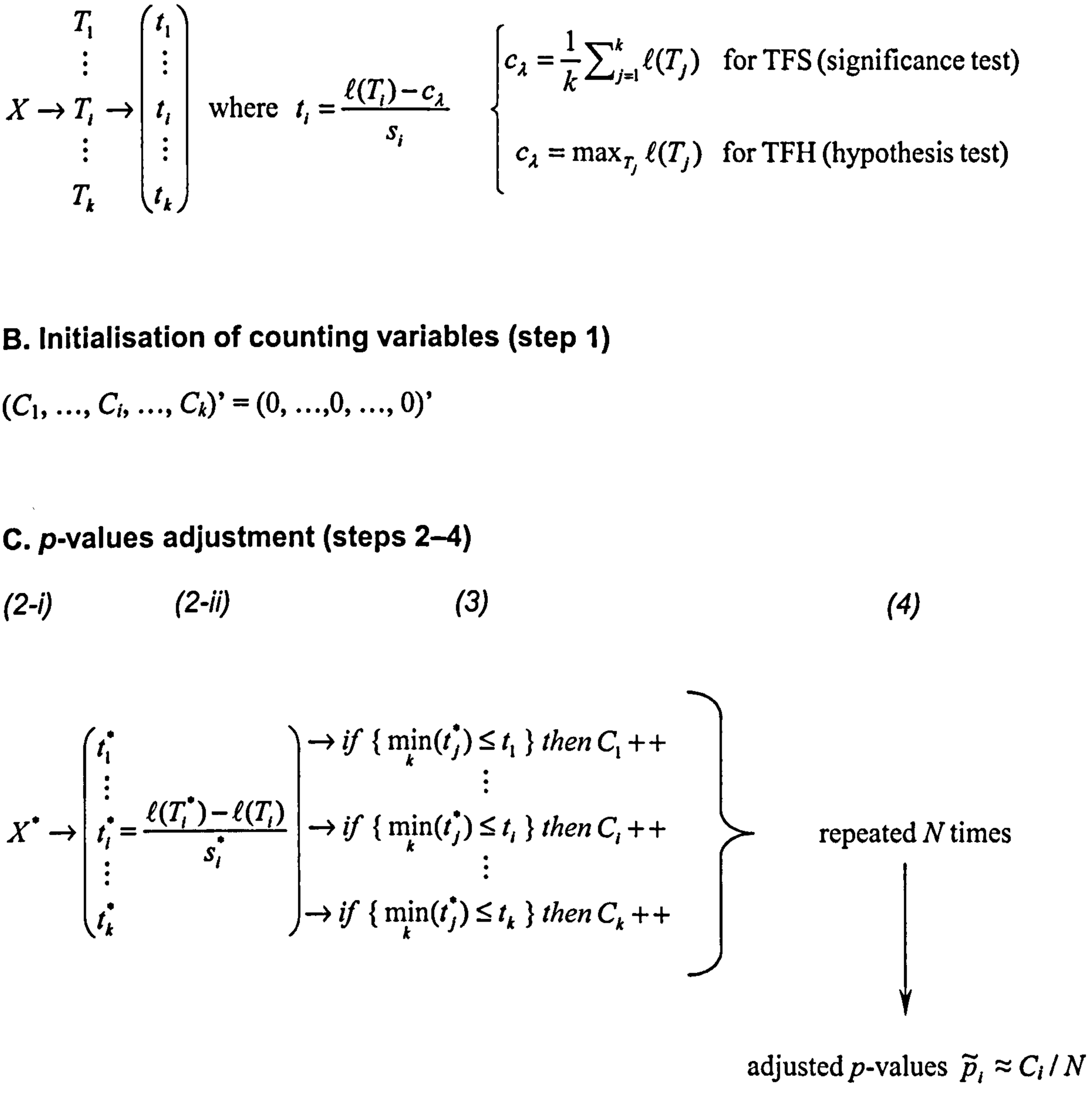


Figure IV.1. Algorithm for computing the adjusted *p*-values either under significance or hypothesis tests. The null hypothesis is characterised by the critical value c_λ. The *p*-value adjustment is performed by bootstrapping: for each replicate *r*, a bootstrap replicate X^* of X is generated (n^* sitewise log-likelihood values are sampled with replacement). For each tree T_i, the statistic t_i^* is computed and a counter C_i is incremented if t_i^* is the smallest resampled *p*-value. This is repeated N times.

branch lengths B so that the general parameterisation of the model is noted $M = \{\theta, T, B\}$ where $\theta = \{\kappa, \alpha, \pi\}$. When we want to evaluate a component of M, such as a specific tree T_i, it is of interest to know the probability of this tree given the data, $p(T_i \mid X)$, that is the posterior probability of tree T_i (Yang and Rannala, 1997; Larget and Simon, 1999). The posterior probability of the complete model of molecular evolution M is obtained by the Bayes theorem as:

$$p(M \mid X) = p(X \mid M)p(M)/p(X) \qquad \text{(IV.4)}$$

The denominator, $p(X)$, is the probability of observing the data X. It can be expressed as $\int p(X \mid M)\, dP(M)$, that is, the likelihood of M averaged over all the real-valued parameters (θ, B) and topologies (T). The cumulative distribution $P(M)$ is chosen to be non-informative. Assuming that the branch lengths and the parameters of the nucleotide substitution model are independent, the posterior probability of a given tree T_i is obtained by integrating over the parameter space under a prespecified substitution model:

$$p(T_i \mid X) = \int p(X \mid T_i, B, \theta)p(B)p(\theta)\, dB\, d\theta \Big/ p(X) \qquad \text{(IV.5)}$$

As noted by Larget and Simon (1999), the normalising constant $p(X)$ is expensive to calculate, as it must also be summed over the tree space. Markov chain Monte Carlo (MCMC) is employed to approximate the posterior distribution of tree T_i (Chapter II), computed as the proportion of trees T_i sampled at stationarity that appear in the posterior MCMC sample.

It is also possible to obtain the posterior probability of tree T_i directly, by using the Bayes theorem for the K possible trees:

$$p(T_i \mid X) = p(X \mid T_i)p(T_i) \Big/ \sum_{j=1}^{K} p(X \mid T_j)p(T_j) \qquad \text{(IV.6)}$$

If the trees are assumed to be equally probable, this reduces to $p(X \mid T_i) \Big/ \sum_{j=1}^{K} p(X \mid T_j)$, which can be approximated by $\exp\{\ell(T_i)\} \Big/ \sum_{j=1}^{K} \exp\{\ell(T_j)\}$ where $\ell(T_i)$ is the log-likelihood of T_i (Kishino and Hasegawa, 1989; Yang and Rannala, 1997; Shimodaira, 2001). This is an approximation because uncertainty about parameters (θ, B) is ignored. As in the frequentist case, this should be computed for the $K = (2s-5)!/\{2^{s-3}(s-3)!\}$ trees of s species. In practice, most of the trees have negligible probabilities, and the computation the approximation is limited to the $k < K$ evaluated trees (Yang and Rannala, 1997), even though it may be imprecise for small values of k. This approximated posterior is noted AP hereafter.

IV.3 – The Bayes approach II: the Bayes factor (BF)

Model selection attempts to identify the (likely) best model. Different measures such as AIC (e.g. Kishino and Hasegawa, 1990b) were presented in Chapter I (see I.3). In this chapter, I use the Bayes factor (BF) as suggested by Yang and Rannala (1997), and present a different implementation from the one recently given by Huelsenbeck and Imennov (2002). Although the BF can be seen as a device to transform posterior probabilities to another comparative scale, the BF has the advantage to represent a test statistic, by means of which statistical decisions can be made. Below, I first present Bayesian hypothesis tests, show how to extend this approach to significance tests in a second subsection, and finally give the computational details to perform these tests.

IV.3.a – Hypothesis test about the ML tree (TBH)

The Bayes factor is defined in terms of a decision rule to compare pairs of models (e.g. Kass and Raftery, 1995). In testing tree T_1 against T_2, this quantity is computed as the ratio of the probabilities of the data under each model:

$$BF_{1,2} = p(X \mid T_1)/p(X \mid T_2) \qquad \text{(IV.7)}$$

This ratio can be seen as the test statistic of a simple hypothesis against a simple hypothesis test (e.g. Bernardo and Smith, 2000, pp. 391–394), although each $p(X \mid T)$ is actually $\int p(X \mid T)\, dB\, d\theta$. If $BF_{1,2}$ is greater than 1, T_1 is favoured by the data against T_2, while BF $\geq$ 10 is considered as strong evidence in favour of T_1 (e.g., Kass and Raftery, 1995). Let us assume that T_1 is the ML tree, T_{ML}. Evaluated against the $k-1$ alternative trees, the object of $BF_{ML,i}$ is to examine whether some tree T_i is as plausible as T_{ML} in light of the available data. Note that $BF_{ML,i}$, like $\tilde{p}_i^{\,ML}$, is not sensitive to the number of tested trees.

IV.3.b – Significance test of trees (TBS)

The use of the Bayes factor as defined by equation (IV.7) can be extended to perform composite vs. simple tests (Bernardo and Smith, 2000, pp. 391–394). To represent adequately the implicit alternative hypothesis characterising significance tests, the composite hypothesis is taken as an average over the parameter space (tree space, including branch lengths B and parameters θ of the substitution model). This leads to an extension of the (pairwise) simple vs. simple test $BF_{ML,i}$ of equation (IV.7):

$$BF_{\bar{T},i} = \int_{T,B,\Theta} p(X \mid T)\, dP(T,B,\theta) \Big/ p(X \mid T_i) \qquad \text{(IV.8)}$$

The numerator is averaged over the tree space T and the spaces B and Θ of real-valued parameters (branch lengths and parameters of the substitution model, respectively). For computational reasons, I use the geometric average of the likelihood integrated over B and Θ. As above, a convenient approximation is to restrict equation (IV.8) to the set of the k tested trees. If moreover, a uniform distribution for the different trees is assumed, it is obtained:

$$BF_{\bar{T},i} = \prod_{j=1}^{k} p(X \mid T_j)^{1/k} \Big/ p(X \mid T_i) \qquad \text{(IV.9)}$$

On a log-scale, equation (IV.9) reduces to an easily computed expression:

$$\log BF_{\bar{T},i} = \frac{1}{k}\sum_{j=1}^{k} \log p(X \mid T_j) - \log p(X \mid T_i) \qquad \text{(IV.10)}$$

As with $BF_{ML,i}$, the composite hypothesis $\bar{T}$ is favoured against T_i when $BF_{\bar{T},i}$ is greater than 10. When it is not the case, T_i is included in the set of trees not significantly worse than any of the tested trees, i.e. in the confidence set of trees at the Bayes equivalent of the lfd.

IV.3.c – Computation

For each tree, $p(X \mid T_i)$ is obtained by integrating over the space Θ of real-valued parameters. Note that unlike in section IV.2, the topology is kept constant in the MCMC used to approximate $p(X \mid T_i)$. The reason behind this is to be able to sample enough points from all the trees included in the analysis, including those with potentially low posterior probability. If we assume that the branch lengths and the parameters of the substitution model are mutually independent, we have:

$$p(X \mid T_i) = \int_{\Theta} p(X \mid T_i, B_i, \theta_i) p(B_i) p(\theta_i)\, dB_i\, d\theta_i \qquad \text{(IV.11)}$$

The index i in the right hand side of equation (IV.11) emphasizes the dependence of the parameters on the model T_i. This integration is not carried out directly: its left hand side is estimated by the harmonic mean of the likelihood with respect to the posterior distribution (Raftery, 1996):

$$p(X \mid T_i) \approx \left(\frac{1}{N} \sum_{t}^{N} 1 / p(X \mid \theta_i^t) \right)^{-1} \qquad \text{(IV.12)}$$

where θ_i^t is sampled from the posterior distribution at the t^{th} step along the MCMC and where N is the number of steps sampled for inference. Details of the implementation can be found in Chapter II (see II.1.e). As required by the use of Bayes factors, only proper prior distributions are used on the parameter space (e.g., see Kass and Raftery, 1995).

Table IV.1. Results of statistical tests of topologies for HIV-1 *gag* and *pol* gene nucleotide data set under two substitution models.

					hypothesis tests		significance tests		
Tree	$\Delta\ell$	pRELL	$p(T_i \mid X)$	AP	$\tilde{p}_i^{\,ML}$	$BF_{ML,i}$	pSH	$\tilde{p}_i^{\,lfd}$	$BF_{\bar{T},i}$
HKY85+Γ									
T_{ML}	0.00	0.792	0.976	0.994	–	–	–	–	**0.014**
T_2	6.03	0.081	0.014	0.002	0.004	$6 \cdot 10^2$	**0.160**	**0.166**	**7.684**
T_3	5.52	0.127	0.010	0.004	0.007	$5 \cdot 10^2$	**0.190**	**0.225**	**9.377**
JC69									
T_{ML}	0.00	0.961	1.000	1.000	–	–	–	–	$\mathbf{9 \cdot 10^{-10}}$
T_2	29.48	0.021	0.000	0.000	0.000	$2 \cdot 10^{13}$	**0.035**	0.000	$5 \cdot 10^4$
T_3	30.54	0.017	0.000	0.000	0.000	$5 \cdot 10^{13}$	**0.027**	0.000	$2 \cdot 10^4$

NOTE. –The tested trees are: T_{ML}: ((B,D),((E2,E1),A2),A1); T_2: ((B,D),(A2,A1),(E2,E1)) and T_3: ((E2,E1),((B,D),A2),A1). $\Delta\ell$ denotes the log-likelihood difference between T_{ML} and T_i. pRELL is the estimate of the bootstrap probability based on the RELL approximation. $p(T_i \mid X)$ is the posterior probability of T_i and AP is its approximation based on equation (IV.6). pSH is the *p*-value of the SH test. $\tilde{p}_i$ denotes the adjusted *p*-values: $\tilde{p}_i^{\,ML}$ for the TFH test and $\tilde{p}_i^{\,lfd}$ for the TFS test. $BF_{\bar{T},i}$ is the Bayes factor of the mean tree vs. each T_i; $BF_{ML,i}$ is the Bayes factor of T_{ML} vs. T_i.

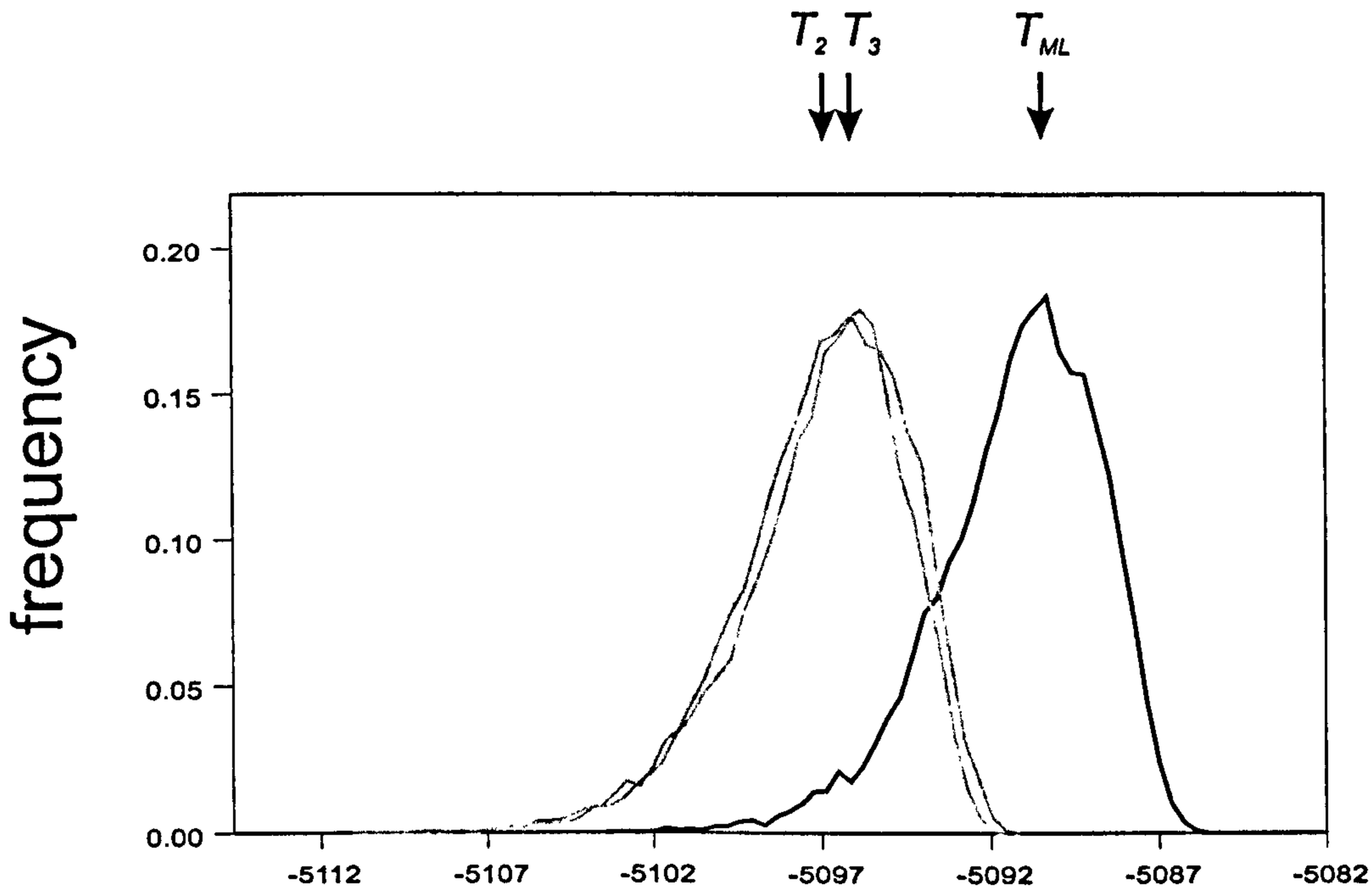
Non-significant results ($\alpha \geq 1\%$; $BF \leq 10$) are in bold.

IV.4 – Application to real data sets

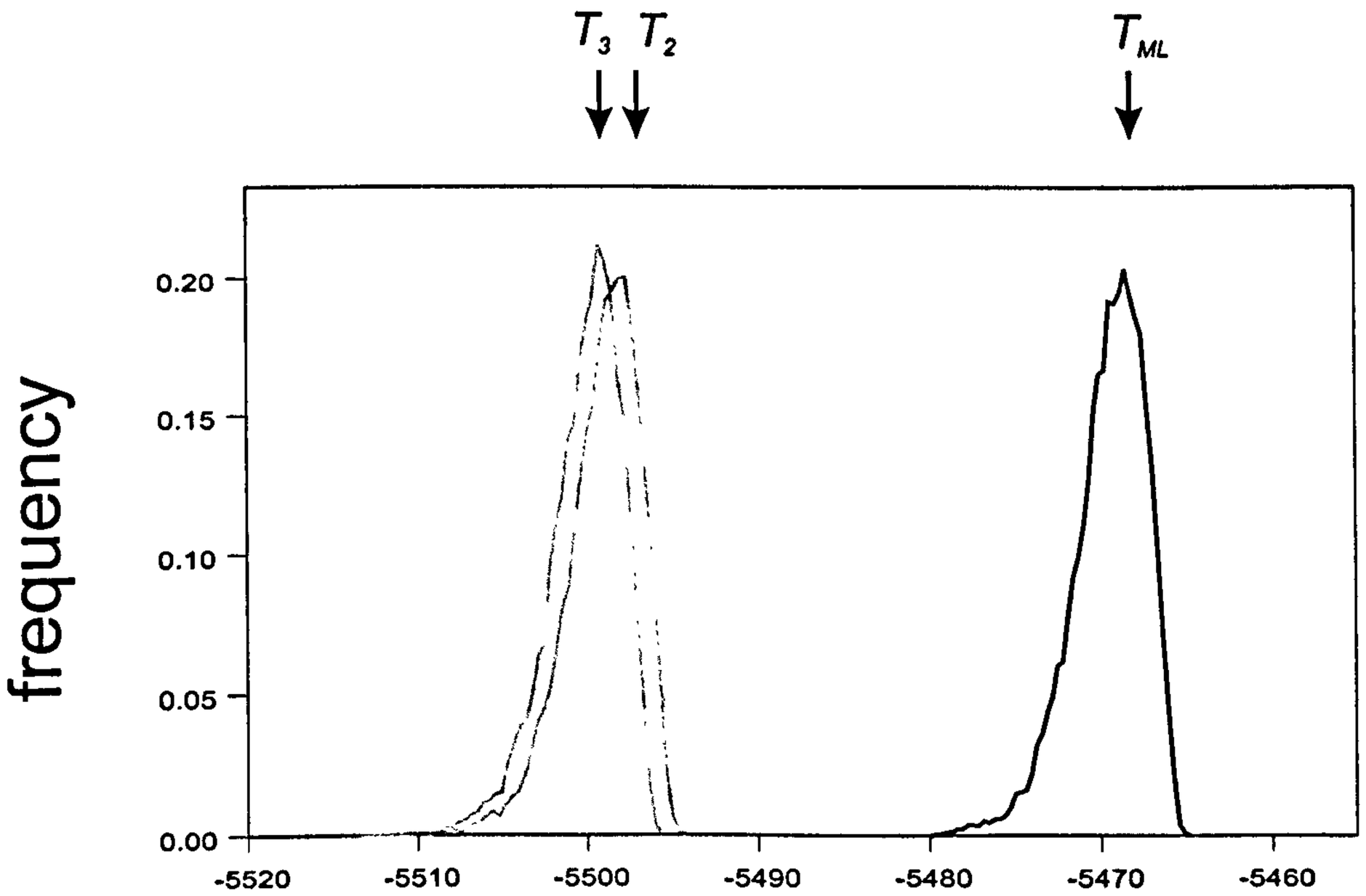
The tests described above are applied to two data sets. All are performed at level 1% in a frequentist approach, or 10 units of likelihood in a Bayes approach. Although there is no formal relationship between these two levels, both are usually interpreted as strong evidence either against (frequentist) or in favour (Bayes) of a hypothesis.

The first data set consists of six HIV-1 nucleotide sequences originally compiled by Goldman et al. (2000): subtypes A (two sequences: A1 and A2), B (one sequence), D (one sequence) and E (two sequences: E1 and E2). These sequences are 2,000 nucleotides long and include the *gag* and *pol* genes. Although the conventional tree groups A1 and A2 together, T_{ML} is ((B,D),((E2,E1),A2),A1) (Goldman et al., 2000). Under the HKY85 + Γ nucleotide substitution model (Hasegawa et al., 1985; Yang, 1994b), the two testing frameworks, hypothesis and significance tests, lead to different *p*-values and to different biological conclusions (Table IV.1). High confidence is put in T_{ML} as being the correct tree (TFH and TBH), which is consistent with results from the SOWH test (see Goldman et al., 2000). On the other hand, the trees T_2 and T_3 are not significantly worse than T_{ML} at the lfd (SH, TFS and TBS; see Figure IV.2A). Although this does not absolutely rule out that, within a testing framework (significance or hypothesis tests), different tests may have different power, it is clear that different approaches (frequentist, either parametric or non-parametric, and Bayes) lead to similar conclusions. Note that the posterior probability of T_{ML}, valued .976, is larger than the bootstrap proportion (pRELL = .792), an effect already mentioned (Yang and Rannala, 1997). The approximation of equation (IV.6), AP (Table IV.1), is larger than the posterior probability computed along the MCMC; however it is likely that the small number of trees compared introduces a bias in this approximation.

A



B



$\log p(X|T_i)$

Figure IV.2. Distribution of the log-likelihood as sampled from the posterior distribution of the tree topologies T_i's for the three HIV-1 trees tested. A vertical arrow indicates the value of the log-probability for each tree, computed as the average of the distribution. The ML tree is in black; all the alternatives are in grey. **A** results under the HKY85 + Γ nucleotide substitution model. **B** results under the JC69 nucleotide substitution model.

As the true substitution model is unknown, I also have evaluated the phylogenies under JC69 (Jukes and Cantor, 1969). When compared with HKY85 + Γ, this model is rejected by a likelihood ratio test ($2\Delta\ell = (-5085.36) - (-5464.93) = 379.57$; $p<.001$), so that this unrealistic model may help us to understand the effect of model misspecification on the tests. Table IV.1 shows that T_2 and T_3 are rejected under both frameworks (hypothesis and significance tests) by all the test procedures: a very high confidence is put in T_{ML} as being the correct tree (TFH and TBH), and it is also the only tree included in the confidence set of trees at the lfd (TFS and TBS). Overconfidence in T_{ML} can also be seen in Figure IV.2B where the distributions of the log-likelihoods sampled from the MCMC do not overlap, and $p(X \mid T_{ML})$ is distinctly the largest. Misspecification of the substitution model results in putting overconfidence in a single tree. Note that while the SH test excludes T_2 and T_3 from the 5% level confidence set at the lfd, it does not exclude them at the 1% level. The significance tests TFS and TBS implemented here appear more sensitive to misspecification of the substitution model than the SH test. However, the difference between the p-values of the SH test and $\tilde{p}^{\,\mathrm{lfd}}$ is partly due to the test statistic used, weighted by the SE's. Indeed, unweighted test statistics give $\tilde{p}^{\,\mathrm{lfd}}$ closer to the p-values of the SH (not shown).

In a larger data set, I analysed the mitochondrial genomes of six mammalian species: human, harbour seal, cow, rabbit, mouse and opossum. These genomes were analysed by Shimodaira and Hasegawa (1999) and by Goldman et al. (2000), at the amino acid level though. I used the alignments of the complete mitochondrial genome (10,806 nucleotides) compiled and analysed by Yoder and Yang (2000). The HKY85 + Γ model of nucleotide substitution is assumed. The topologies considered are the fifteen selected by Shimodaira and Hasegawa (1999), and are labelled by decreasing order of $\ell(T_i)$ (Table IV.2). This ordering differs slightly from the one found in the aforementioned paper,

Table IV.2. Results of statistical tests of topologies for mammalian whole mitochondrial DNA data set under the HKY85 + Γ model of nucleotide substitution.

					hypothesis tests		significance tests		
Tree	$\Delta\ell$	pRELL	$p(T_i \mid X)$	AP	$\tilde{p}_i^{\,ML}$	$BF_{ML,i}$	pSH	$\tilde{p}_i^{\,lfd}$	$BF_{\bar{T},i}$
T_{ML}	0.00	0.502	0.554	0.579	–	–	–	–	$\mathbf{6{\cdot}10^{-16}}$
T_2	0.32	0.488	0.446	0.421	**0.359**	**2.565**	**0.844**	**1.000**	$\mathbf{1{\cdot}10^{-15}}$
T_3	14.11	0.003	0.000	0.000	0.000	$7{\cdot}10^{6}$	**0.344**	**1.000**	$\mathbf{4{\cdot}10^{-9}}$
T_4	24.96	0.006	0.000	0.000	0.000	$3{\cdot}10^{10}$	**0.093**	**0.999**	$\mathbf{2{\cdot}10^{-5}}$
T_5	32.65	0.000	0.000	0.000	0.000	$5{\cdot}10^{14}$	**0.019**	**0.991**	**0.252**
T_6	34.44	0.000	0.000	0.000	0.000	$5{\cdot}10^{14}$	**0.020**	**0.771**	**0.305**
T_7	39.09	0.000	0.000	0.000	0.000	$2{\cdot}10^{17}$	0.007	0.000	$1{\cdot}10^{2}$
T_8	42.02	0.000	0.000	0.000	0.000	$2{\cdot}10^{18}$	0.003	0.000	$8{\cdot}10^{2}$
T_9	42.81	0.000	0.000	0.000	0.000	$3{\cdot}10^{18}$	0.002	0.000	$2{\cdot}10^{3}$
T_{10}	45.31	0.000	0.000	0.000	0.000	$2{\cdot}10^{19}$	0.001	0.000	$1{\cdot}10^{4}$
T_{11}	46.26	0.000	0.000	0.000	0.000	$4{\cdot}10^{20}$	0.000	0.000	$2{\cdot}10^{5}$
T_{12}	46.27	0.000	0.000	0.000	0.000	$1{\cdot}10^{20}$	0.000	0.000	$6{\cdot}10^{4}$
T_{13}	49.20	0.000	0.000	0.000	0.000	$7{\cdot}10^{20}$	0.001	0.000	$4{\cdot}10^{5}$
T_{14}	51.14	0.000	0.000	0.000	0.000	$9{\cdot}10^{21}$	0.000	0.000	$5{\cdot}10^{6}$
T_{15}	56.38	0.000	0.000	0.000	0.000	$6{\cdot}10^{24}$	0.000	0.000	$3{\cdot}10^{9}$

NOTE. –The tested trees are: T_{ML}: (((H,(P,B)),O),M,D), T_2: (((H,O),(P,B)),M,D), T_3: ((H,((P,B),O)),M,D), T_4: (((H,(P,B)),M),O,D), T_5: ((H,(P,B)),(O,M),D), T_6: (((H,O),M),(P,B),D), T_7: ((H,O),((P,B)M),D), T_8: (((H,M),(P,B)),O,D), T_9: ((H,((P,B),M)),O,D), T_{10}: ((H,(O,M)),(P,B),D), T_{11}: ((H,M),((P,B),O),D), T_{12}: (H,(((P,B),O),M),D),

T_{13}: (((H,M),O),(P,B),D), T_{14}: (H,((P,B),(O,M)),D) and T_{15}: (H,(((P,B),M),O),D) where the taxa are labelled as: H = *Homo sapiens* (human), P = *Phoca vitulina* (harbour seal), B = *Bos taurus* (cow), O =*Oryctolagus cuniculus* (rabbit), M = *Mus musculus* (mouse) and D = *Didelphis virginiana* (opossum). $\Delta\ell$ denotes the log-likelihood difference between T_{ML} and T_i. pRELL is the estimate of the bootstrap probability based on the RELL approximation. $p(T_i \mid X)$ is the posterior probability of T_i and AP is its approximation based on equation (IV.6). pSH is the *p*-value of the SH test. $\tilde{p}_i$ denotes the adjusted *p*-values: $\tilde{p}_i^{\,ML}$ for the TFH test and $\tilde{p}_i^{\,lfd}$ for the TFS test. $BF_{\bar{T},i}$ is the Bayes factor of the mean tree vs. each T_i. $BF_{ML,i}$ is the Bayes factor of T_{ML} vs. T_i. Non-significant results ($\alpha \geq 1\%$; $BF \leq 10$) are in bold.

where the likelihood values were computed from amino acid data; the ordering of the trees is however identical when only codon positions one and two were used. Two trees (T_{ML} and T_2) are strongly supported by RELL, $p(T_i \mid X)$ or its approximation AP, with probabilities close to ½. Likewise, hypothesis tests (TFH and TBH) do not exclude T_2 when testing for the correct tree (Table IV.2 and Figure IV.3), whereas all the other trees are excluded. Note also that T_{ML} and T_2 could not be distinguished by confidence measures such as RELL, $p(T_i \mid X)$ or its approximation AP.

From a significance test perspective, the confidence set of trees at the lfd is larger than T_{ML} and T_2. Whichever significance test is used, trees T_3 to T_6 are included (Figure IV.3). While TFS estimates the same confidence set as TBS and SH, the *p*-values of TFS ($\tilde{p}_i^{\,\mathrm{lfd}}$) are generally larger than those of the SH test, especially for T_5 and T_6. Again, this effect is partly due to weighting the test statistic used for computing $\tilde{p}_i^{\,\mathrm{lfd}}$. In this significance framework, the traditional Glires grouping (Lagomorpha + Rodentia), represented by trees T_5, T_{14} and T_{10}, has little support when the three codon positions are considered. The Primates + Lagomorpha grouping has more support since T_2 and T_6 (but not T_7) are included in both confidence sets. Results at the nucleotide level appear inconsistent with those at the amino acid level obtained by Shimodaira and Hasegawa (1999). When only the first two codon positions are used, the trees T_{ML} to T_6 and T_{10} to T_{14} are included in the confidence set at level 1%. This corresponds to the results obtained by Shimodaira and Hasegawa (1999), which suggests that the results obtained with all three codon positions are due to the effect of saturation.

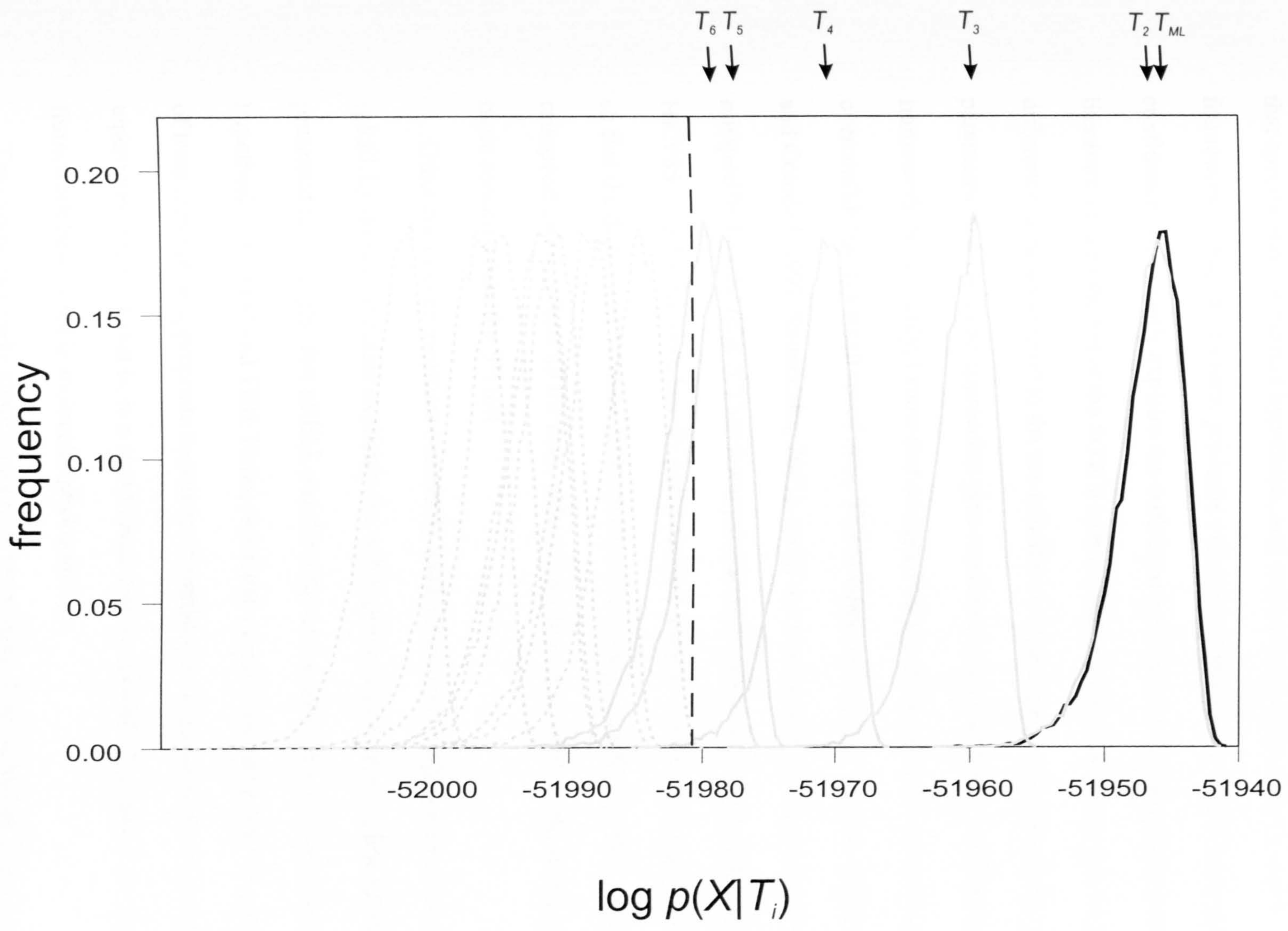


Figure IV.3. Distribution of the log-likelihood as sampled from the posterior distribution of the tree topologies T_i for the fifteen mammalian trees tested. The ML tree is in black; all the alternatives are in grey. The vertical broken line indicates the average probability of the data over the fifteen tested trees (Bayesian equivalent of the lfd), on the right of which are the trees included in the confidence set (T_{ML} to T_6, identified by vertical arrows); grey distributions in a broken line are not included in the set.

IV.5 – Discussion and conclusions

By introducing four new tests of molecular phylogenies, this chapter principally shows that specification of the null hypothesis, and of an alternative hypothesis in the framework of hypothesis tests, produces a dramatic difference in terms of measures of confidence. This mostly explains the difference between tests previously described in the literature such as the SH or the SOWH tests (Goldman et al., 2000). In particular, this difference does not amount to the use of different techniques, since parametric, non-parametric and Bayesian approaches give consistent results within a given testing framework. Secondarily, I show that misspecification of the substitution model gives overconfidence in a small set of trees. This is supported by other studies (Huelsenbeck and Crandall, 1997; Shimodaira, 2001), and is not unexpected as in this case, the misspecified model (e.g. JC) has fewer parameters than the selected model (e.g. HKY85 + Γ). Under more heavily parameterised models, the likelihood surface is flatter, so that the distinction between different trees becomes more difficult. This sensitivity to misspecification affects all the tests, although the tests presented in this study appear more sensitive than the SH test.

Other confidence measures commonly used, such as the bootstrap proportion of trees (pRELL), do not formulate explicitly the null hypothesis they rely on. The results presented here suggest that pRELL and the posterior probability of a tree behave like the hypothesis tests TFH and TBH. Hence, bootstrap proportions and posterior probabilities of trees may not be appropriate measures of confidence when the objective is to compare topologies at the lfd, that is, in a significance test framework. This raises the issue of the framework best suited in molecular phylogenetics.

Two properties make significance tests interesting. First, unlike hypothesis tests, they explicitly work on a set of trees, hereby avoiding repeated pairwise comparisons against

the maximum likelihood tree. Pairwise comparisons typically lead to suboptimal procedures, in the sense that they do not use all the available information from the data. A similar situation is encountered when distance methods are used to estimate phylogenetic trees (Swofford et al., 1996) or when counting methods are used to detect positive selection (Yang and Nielsen, 1998). Second, significance tests can be interpreted as misspecification tests. Given the null hypothesis H_0 of a significance test, its implicit alternative is the complement of H_0, not only with respect to a specific tree as in hypothesis tests, but more generally with respect to the model space (the set of all probability measures). Therefore, low significance, as measured by the *p*-value associated with a given tree, indicates that the null model poorly describes the data at hand. This either means that a binary tree topology (or a specific set of such trees) is misspecified, or means that the substitution model is grossly wrong, or that both tree and substitution model are poor descriptors.

Choosing a good test generally amounts to selecting a procedure that maximises the power, that is the probability of rejecting the null hypothesis H_0 when it is false, subject to the condition that the probability of rejecting H_0 when it is true is less than or equal to α. In the *p*-value adjustment procedure used here, the FEW is corrected at level α, at least approximately (Westfall and Young, 1993, p. 53). But a test where only the FWE is controlled may not be the most powerful. At the moment, it is not clear whether TFS or TBS are more powerful than the SH test, but all are, by construction, not uniformly powerful, as significance is dependent on the set of the tested trees. This problem, noted by Goldman et al. (2000) in the case of the SH test, may appear as a weakness compared to hypothesis tests, which are immune to this effect.

The main objective of this chapter was not to propose computationally efficient tests, but to explain the difference between the SH and the SOWH test. Future efforts should

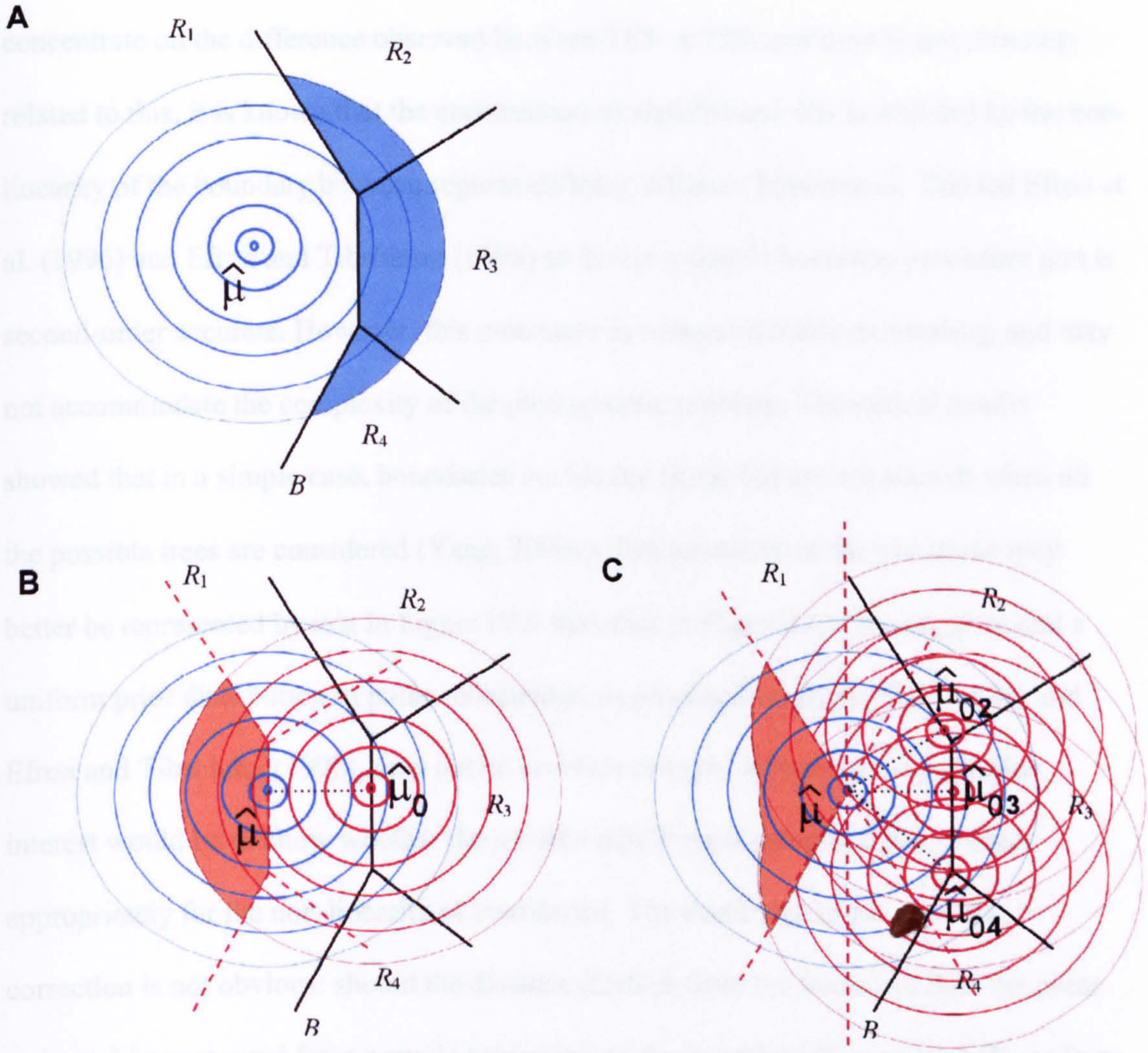


Figure IV.4. A working hypothesis: does the procedure adjusting *p*-values for multiple testing also correct for curved boundaries and how is it done? **A**. rejection level under the naïve bootstrap. **B**. construction of the rejection level for a single hypothesis. **C**. rejection level for multiple hypotheses. (see text and Figure I.4 for details.)

concentrate on the difference observed between TFS or TBS and the SH test. Possibly related to this, it is known that the construction of significance sets is affected by the non-linearity of the boundary between regions defining different hypotheses. This led Efron et al. (1996) and Efron and Tibshirani (1998) to devise a double bootstrap procedure that is second-order accurate. However, this procedure is computationally demanding, and may not accommodate the complexity of the phylogenetic problem. Theoretical results showed that in a simple case, boundaries are locally linear but are not smooth when all the possible trees are considered (Yang, 2000a). The geometry of the tree space may better be represented by that in Figure IV.4 than that in Figure I.4. This suggests that a uniform prior distribution in polar coordinates, as proposed by Efron et al. (1996) and Efron and Tibshirani (1998), may not be accurate enough. A question of particular interest would be to know whether the *p*-value adjustments presented here correct appropriately for the non-linearity of boundaries. The exact way to perform this correction is not obvious: should the distance (further from the boundary than the point estimate) be measured from a single projection on the boundary (Figure IV.4 **B**), or from the projection on each boundary (Figure IV.4 **C**)? Note however that the area representing the rejection region under the naïve bootstrap (Figure IV.4 **A**, blue area) is larger than any of the other areas (Figure IV.4 **B**, **C**, orange area). This can be compared with the results obtained from the two data sets considered above, where $\tilde{p}_i^{\,ML} <$ pRELL (Tables IV.1 and 2).

To conclude, the SH, SOWH and the four new tests appear sensitive to misspecification of the substitution model. The difference observed between tests of phylogenetic hypotheses amounts to the testing framework used. Hypothesis tests evaluate the null hypothesis that a given tree is correct and formulate explicitly an alternative hypothesis, usually taken at the ML tree. Significance tests formulate a null

model (set of trees) whose significance indicates how satisfactory this model is given the observed data. Bayesian formulations of these testing frameworks give consistent results, showing that the techniques used are not responsible for the difference between the SH test and the SOWH test.

Chapter V

Model averaging, multiple genes and detecting positive selection

V.1 –Bayesian model averaging: a response to model misspecification?

Choosing a single model can seriously affect the estimate and leads to underestimating the variance. Model averaging can then be the solution to this estimation problem. MCMC Model Composition (MC^3, see Madigan and York, 1995) uses a stochastic chain through model space that has an equilibrium distribution equal to the true posterior model distribution. Applied to discrete graphical models in the context of estimating a closed population estimate, these authors report an improved predictive performance. It is exactly in this framework that tree selection procedures presented in Chapter IV were implemented, although I did not introduce it this way. George and McCulloch (1993) have developed the stochastic search variable selection method (SVSS), which is similar to MC^3, but models the parameters as a mixture of normal distributions, with components of small and large variance. For "non-relevant" variables, the component with small variance will have large probability, effectively "selecting out" that variable. Their framework represents a traditional hierarchical model, and they can run a standard Gibbs or a Metropolis-Hastings sampler.

When model selection is not the issue, the standard Bayesian solution to the problem of model uncertainty is to use a set of models, instead of just one model, for making both inference and prediction. Let θ be any quantity of interest such as divergence times. If $\mathbf{M} = \{M_1; \ldots; M_K\}$ denotes the set of all models considered, then the posterior distribution of θ given the data X is:

$$p(\theta \mid X) = \sum_{k=1}^{K} p(\theta \mid M_k, X)\, p(M_k \mid X) \qquad \text{(V.1)}$$

where $p(\theta \mid M_k, X)$ is called the predictive distribution of the quantity of interest, and $p(M_k \mid X)$ is the posterior model probability for model M_k. Therefore, equation (V.1) is a weighted average of the predictive distributions with respect to the different models. The weights correspond to the posterior model probabilities.

It is possible to apply these ideas to the estimation of divergence times, where one of the possible criticisms is that the topology of the tree is fixed. The effect and the extent of such a model uncertainty are not understood. The solution would be to integrate over the tree space, in the same way as it is done in Chapter IV, along with integration over divergence times and rates. The cost of such a model would be an increase in terms of memory requirements (all the sampled trees with their vectors of times and rates would have to be kept) and running time. A less expensive approach may be to select a few competing topologies, and to average only over these models (Madigan and Raftery, 1994). The selection should ideally be based on the approach followed in Chapter IV, to obtain a set of candidate trees (at the lfd), or at least those with the largest posterior probability. When the data have not enough information to support a small number of trees, only the models belonging to set A defined below are considered.

$$A = \left\{ M_k \setminus \frac{\max_l \{p(M_l \mid X)\}}{p(M_k \mid X)} \leq C \right\} \qquad \text{(V.2)}$$

Madigan and Raftery (1994) show that a constant $C \approx 20$ provides a good approximation to averaging over the whole model space, at least for a certain class of models.

This approach is usually interesting in cases where different models lead to different estimates. In view of the results in Chapters II and III, different models of rate change lead to close date estimates, at least with real data (but see section II.3). In the same line, different substitution models usually give different estimates. But in this case it is not sure whether Bayes averaging methods would make sense, while MC^3 probably would.

Model uncertainty is one component of model misspecification. Averaging over approximating models has the merit of adjusting the variance of the final estimate. But it is likely that the effect of misspecification of the substitution model may require the

implementation either of more complex (realistic) models, or the development of tests less sensitive to misspecification (see Golden, 1995).

V.2 – Multiple genes models

The models described in Chapters II and IV are for a single gene. In Chapter III, I have analysed several genes, assuming they are *a priori* independent. This may not be the case, and the evolution of some of them may be correlated. In particular, it is sensible to think that evolution of the mitochondrial genes exhibits the same trends, as this set of genes actually constitutes a single locus. A natural extension of the model is to analyse data from multiple genes, taking into account possible correlation between genes. Far from being exhaustive, I will here present three possible extensions of the model to estimate divergence dates under models of rate change.

V.2.a – General model and multivariate-normal approximation

The most general model would consider the joint posterior probability $p(X \mid \theta)$, where both X and θ are vector valued. Restricting the model to two genes, it can be written as:

$$p(\theta \mid X_1, X_2) = \frac{p(X_1, X_2 \mid \theta)\, p(\theta)}{p(X_1, X_2)} \qquad \text{(V.3)}$$

The difference with the theory developed in Chapter III is relative to the handling of the likelihood of θ, i.e. $p(X_1, X_2 \mid \theta)$. As the joint likelihood function for two genes may be difficult to compute, it is however possible to use asymptotic results. Note that under a fixed topology the likelihood functions for X_1 and X_2 are the same. Under some regularity condition about the two probability density functions $p(x_1 \mid \theta_1)$ and $p(x_2 \mid \theta_2)$, the maximum likelihood estimator of θ is asymptotically distributed with mean the vector of marginal means $\theta = \{\theta_1, \theta_2\}$ and variances $\sigma_i = -E_\theta\left\{\frac{\partial^2}{\partial \theta_i^{\,2}} \log p(X_i \mid \theta_i)\right\} / n\Delta$, where

$$\Delta = E_\theta\left\{\frac{\partial^2}{\partial\theta_1^{\,2}}\log p(X_1\,|\,\theta)\right\}E_\theta\left\{\frac{\partial^2}{\partial\theta_2^{\,2}}\log p(X_2\,|\,\theta)\right\} - \left(E_\theta\left\{\frac{\partial^2}{\partial\theta_1\,\partial\theta_2}\log p(X_1, X_2\,|\,\theta)\right\}\right)^2.$$

The correlation ρ can be approximated by the expectation

$$E_\theta\left\{\frac{\partial^2}{\partial\theta_1\,\partial\theta_2}\log p(X_1, X_2\,|\,\theta)\right\}$$ (e.g. Mood et al., 1974, pp. 360-361).

V.2.b – Gene partitioning

A simplifying perspective is to view X_1 and X_2 as two partitions of the same data set. The underlying assumption is that the two genes are co-segregating, and hence have the same phylogeny. Divergence times are thus constrained to be the same across the different partitions. If the two genes are strictly correlated, their branch lengths are proportional, as in the sense of the model developed by Yang (1996b).

The interesting point is that θ contains only one vector of divergence times. As this model assumes that all the gene partitions reflect the same history, the different genes share the same divergence times. Alongside of this constraint, incorporating fossil-derived upper and lower bounds on divergence dates in the algorithm should improve convergence.

V.2.c – Independent rates

Because the former model assumes that branch lengths are proportional across different genes, the implicit assumption is that rates are also proportional among different gene partitions. This assumption is somewhat unrealistic, as it is likely that genes in different parts of the genome, such as different chromosomes, are likely to evolve independently. This is all the more true as different selective pressures are expected to affect different genes.

To circumvent this last issue, a simple approach would be to assume that different genes only share divergence times, with their respective rates of molecular evolution evolving independently on the rooted phylogenetic tree. This is similar to the model described in V.2.a, but the absence of correlation between genes makes it possible to approach the problem without having to resort to a normal approximation (since the requirement to model interaction disappears). From there, two possibilities exist for the treatment of rate autocorrelation (hyperparameters of the prior distributions for each gene): either these hyperparameters are not shared among genes, hereby allowing some genes to have evolved in a more "clocklike" manner than others, or the different genes may share these hyperparameters. The first option is more general, but is also more heavily parameterised. These models treating multiple genes should deserve more focus in the future, as they are likely to emulate the results presented in Chapter III.

V.3 – Detecting sites under positive selection

Changes at the amino acid level should reflect the selective pressures acting on proteins. With respect to amino acids, two types of nucleotide changes can occur. After a mutation, the amino acid type encoded by a mutated codon can be modified. This first type of mutation is called non-synonymous. Because of the structure of the genetic code (degeneracy), the amino acid of a mutated codon can remain unchanged. This situation is referred to as a synonymous mutation. After the action of selective pressures, i.e. after a certain length of evolutionary time, substitutions are observed. When a site (codon) is neutral, it is expected that non-synonymous substitutions are as frequent as synonymous changes. However, a "deficit" of non-synonymous substitutions at a codon position indicates that selective pressures have acted on this site presumably because it plays a key role in the functioning of final protein. This site is said to be under negative selection,

which, in functionally important proteins, is expected to be the rule. Conversely, detecting more non-synonymous than synonymous substitutions suggests the action of positive or diversifying selection. A convenient measure of selection appears as the ratio of the rates of non-synonymous to synonymous substitutions. This ratio, traditionally denoted d_N/d_S, or ω (e.g. Yang, 2001), can be used to identify regions of a protein that are under negative selection or, more interestingly, under positive selection.

Originally, identifying methods were counting the number of differences between pairs of sequences (Miyata and Yasunaga, 1980; Li et al., 1985; Nei and Gojobori, 1986). This crude method was then improved to take multiple changes at the nucleotide level into account (Li, 1993; Tajima, 1993b; Ina, 1995; Yang and Nielsen, 2000), but the technique remains imprecise as it averages over all the sites of a sequence. An improvement to these *ad hoc* methods comes from modelling the evolution of the ω ratio to account for its variation over lineages (Yang, 1998; Yang and Nielsen, 1998) or among amino acid sites (Nielsen and Yang, 1998; Yang et al., 2000b).

The general Markov model can be found in Goldman and Yang (1994). Among-site (codon) variation of ω can be modelled by partitioning the data, when some prior information such as the location of structural and functional domains is available (Yang, 2001; Yang and Swanson, 2001). Otherwise, a prior distribution is assumed for the ratio variation (Nielsen and Yang, 1998; Yang et al., 2000a), which leads to the same treatment as for among-site rate variation (see section I.1.c). In the simplest case, a number K of ω ratios categories is assumed and the likelihood function appears as a mixture of pdf's whose proportions are estimated (ML) jointly with the average ratio for each class. Sites belonging to a given category are identified by an empirical Bayes approach according to their posterior probability.

Alternatively, a full Bayes approach can be undertaken. The distribution of interest is $p(\omega_i \mid X)$, where ω_i is for site *i*. From the Bayes theorem we have $p(\omega_i \mid X) = p(X \mid \omega_i)$ $p(\omega_i) / p(X)$, which is the marginal probability $\sum_{j \neq i} p(\omega_j \mid X)$, considering that sites are iid. The prior distribution $p(\omega_i)$ is uniform. The disadvantage of such a model is that it is heavily parameterised – to the point of non-identifiability (there are more parameters to estimate than there are data points). As above, sites can be grouped into a finite number of categories, or a continuous prior, conveniently a mixture of (log) normal distributions, can be chosen.

Across lineage ratio variation can be added. A possibility would be to model the evolution of the ω_i ratio with an autoregressive process such as the one used to model rate variation at the nucleotide level to estimate divergence dates (Chapter II). A somewhat less heavily parameterised option would be to take a discrete approach to model non-synonymous and synonymous mutations along the tree using a compound Poisson process as implemented in Huelsenbeck et al. (2000). More recently, a mutation mapping procedure was proposed (Nielsen and Huelsenbeck, 2002), where the expected number of non-synonymous substitutions at a site given the data $p(d_N \mid X)$ is averaged over all possible configurations (mappings). This probability, $\int_\Theta p(d_N \mid X, \theta)\, p(\theta \mid X)\, d\theta$, is itself approximated using an MCMC.

One of the advantages of the Bayesian approach outlined here is to be able to test a large range of hypotheses which are difficult to assess otherwise. This is the case when testing the fit of a gamma distribution to model among-site rate variation (section I.3.a), or testing a discrete model of selection against the neutral case (M2 vs. M1 in the terminology of Yang et al., 2000b). In both instances, the issue of having a parameter at the boundary disappears when adopting a Bayesian approach as the fit is measured by the Bayes factor:

$$B_{M2/M1} = \frac{P(X \mid M2)}{P(X \mid M1)} = \frac{P(X \mid p_1, p_2, \omega_2)}{P(X \mid p_1, p_2 = 0, \omega_2 = 0)} \qquad \text{(V.4)}$$

where the p_i's are the frequencies of the ω_i's rate categories. The fit of more complex models could also be tested, such as a direct comparison of a finite mixture of k lognormal distributions ($\omega \geq 0$) vs. a mixture of beta and gamma distribution (M9 in Yang et al., 2000b), although the models are not nested. The general formulation of this composite vs. composite test (Bernardo and Smith, 2000, p. 392) can be summarised by the following Bayes factor:

$$B_{2/1} = \int_{\Theta_2} P(X \mid \theta)\, P(\theta)\, d\theta \Big/ \int_{\Theta_1} P(X \mid \theta)\, P(\theta)\, d\theta \qquad \text{(V.5)}$$

where a model (2) is compared to (1). Each model is defined by a family of parametric hypotheses on subspaces Θ_i which can be nested, overlapping or strictly non-nested, and over which the different parameters are integrated. In the example taken above we would have the proportions of the mixture and the parameters (mean $\overline{\omega}_i$ and variance $\sigma^2(\omega_i)$) of each lognormal distribution for Θ_2 ($\Theta_2 = \{p_i, \overline{\omega}_i, \sigma^2(\omega_i)\}_{i \in [1,k]}$, $\sum_i p_i = 1$, $\overline{\omega}_i \geq 0$, $\sigma^2(\omega_i) \geq 0$), and Θ_1 would be the mixing proportion of the beta and the gamma distributions plus their respective parameters ($\Theta_1 = \{p_0, p, q, \alpha, \beta\}$).

Conclusions

Chapter I showed that two of the issues emerging from the remark by Sokal and Sneath (1963) quoted in the introduction were still generating a heated debate. More precisely, the estimation of divergence times and rates of evolution involves complicated models, where the parameters are not always identifiable in the traditional ML approach. The approach undertaken in Chapter II showed that modelling rate change increases dramatically the fit when compared to the clock assumption. Irrespective of their complexity level, different models of rate change produce very similar date estimates, closer to what is expected from the fossil records (Chapter III). The second issue is that of testing hypotheses in non-regular cases, where hypotheses are not nested, a posteriori selected, involving multiple comparisons, and defining regions in the parameter space with rugged boundaries. Although the use of the posterior Bayes factor (Chapter II) can be contentious, Chapter IV showed that the traditional Bayes factor can be used to select models such as tree topologies in an intuitive framework. Moreover, it was shown that Bayesian model selection is equivalent to its frequentist counterpart, when the distinction between hypothesis tests and significance tests is correctly made. The Bayes approach examined is easily extended to more complicated models as shown in Chapter V, which makes of it a promising and powerful framework for future studies.

More generally, the power of the Bayesian approach relies on the interplay of two factors. First, the possibility to develop complicated models makes a non-trivial difference when all the available information can be used. Objectivity is often penalised, but the fit to the data is substantially improved. Bayesian modelling would not be of much use without the second component: the use of MCMC techniques to sample efficiently from target distributions with no closed-form expressions. It is this interplay between Bayesian modelling and Bayesian computation that makes the approach so powerful. This power encompasses three aspects. Unlike ML analysis, computational

costs are not multiplicative in the Bayesian approach. For instance, searching the tree space under traditional approaches involves greedy re-optimisation steps of all the parameters for each topology examined. The Bayesian approach avoids this problem by integrating over all the parameters while searching the tree space. Second, the Bayesian approach gives an intuitive measure of confidence, the credible set, so there is no need to resort to parametric or non-parametric tests. This is of particular interest when uniform prior distributions are used, as in this case credible sets are identical to confidence sets. Third, uncertainty with respect to "nuisance parameters" is integrated out of the model. The Bayesian approach may have a greater power than ML in such situations.

While the use of Bayesian methods in phylogenetics is becoming popular (Huelsenbeck et al., 2001), some questions are left open though. In particular the sensitivity of confidence measures to both model specification (see Appendix 3) and geometry of the tree space calls for the further development of "New Tests" (Perlman and Wu, 1999), including in Bayesian statistics.

Finally, I would like to come back on one advantage of the Bayesian approach in molecular phylogenetics: its ability to manipulate fairly complex models. To quote Geyer (1999), "*If you can write down a model, I can do the likelihood inference for it, not only maximum likelihood estimation, but also likelihood ratio tests, likelihood-based confidence regions, profile likelihoods, whatever. That includes conditional likelihood inference and inference with missing data*". Of course, this is an overstatement, as models can become so complicated that MCMC computation would require years. "*But analyses that can be done are far beyond what is generally recognized*" (Geyer, 1999). The recent increase in the number of publications using this approach shows the need for fast and intuitive methods able to deal with complicated models... along with the appropriate computer programs.

References

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, and Inoko H (2002) Evidence of en bloc duplication in vertebrate genomes. *Nature Genetics* **31**:100-105.

Adachi J and Hasegawa M (1992) Amino acid substitution of proteins coded for in mitochondrial DNA during mammalian evolution. *Japanese Journal of Genetics* **67**:187-197.

Adachi J and Hasegawa M (1996) *MOLPHY: programs for molecular phylogenetics.* Tokyo: Institute of Statistical Mathematics.

Aitkin M (1991) Posterior Bayes factor. *Journal of the Royal Statistical Society (B)* **53**:111-142.

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In Petrox B, Caski F (eds) *Second International Symposium on Information Theory.*

Altman DG and Andersen PK (1989) Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medecine* **8**:771-783.

Ayala FJ, Rzhetsky A, and Ayala FJ (1998) Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. *Proceedings of the National Academy of Sciences (USA)* **95**:606-611.

Bar-Hen A and Kishino H (2000) Comparing the likelihood functions of phylogenetic trees. *Annals of the Institute of Statistical Mathematics* **52**:43-56.

Barry D and Hartigan JA (1987a) Asynchronous distance between homologous DNA sequences. *Biometrics* **43**:261-276.

Barry D and Hartigan JA (1987b) Statistical analysis of hominoid molecular evolution. *Statistical Science* **2**:191-207.

Baumiller TK (1999) Enough remains to work with. *Science* **283**:1271.

Benton MJ (1993) *The fossil record 2.* Benton MJ (ed). London: Chapman & Hall.

Benton MJ (1999) Early origins of modern birds and mammals: molecules vs. morphology. *Bioessays* **21** :1043-1051.

Bernardo JM and Smith AFM (2000) *Bayesian theory.* New York: Wiley.

Bickel DR (2000) Implications of fluctuations in substitution rates: impact on the uncertainty of branch lengths and on relative-rate tests. *Journal of Molecular Evolution* **50**:381-390.

Bowring SA, Grotzinger JP, Isachsen CE, Knoll AH, Pelechaty SM, and Kolosov P (1993) Calibrating rates of early Cambrian evolution. *Science* **261**:1293-1298.

Brasier MD (1998) From deep time to late arrival. *Nature* **395**:547-548.

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**:1393-1398.

Bromham L, Penny D, Rambaut A, and Hendy MD (2000) The power of relative rates tests depends on the data. *Journal of Molecular Evolution* **50**:296-301.

Bromham L, Rambaut A, Fortey R, Cooper A, and Penny D (1998) Testing the Cambrian explosion hypothesis by using a molecular dating technique. *Proceedings of the National Academy of Sciences (USA)* **95**:12386-12389.

Bromham LD and Hendy MD (2000) Can fast early rates reconcile molecular dates with the Cambrian explosion? *Proceedings of the Royal Society London B Biological Sciences* **267**:1041-1047.

Brown RH, Richardson M, Boulter D, Ramshaw JA, and Jefferies RP (1972) The amino acid sequence of cytochrome c from Helix aspersa Muller (garden snail). *Biochemical Journal* **128**:971-974.

Bush RM, Bender CA, Subbarao K, Cox NJ, and Fitch WM (1999) Predicting the evolution of human influenza A. *Science* **286**:1921-1925.

Buss LW and Seilacher A (1994) The phylum Vendobionta: a sister group of the Eumetazoa? *Paleobiology.* **20**:1-4.

Cao Y, Fujiwara M, Nikaido M, Okada N, and Hasegawa M (2000) Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. *Gene* **259**:149-158.

Cavalli-Sforza L and Edwards AW (1964) Analysis of human evolution. *Proceedings of the XI International Congress of Genetics* **3**:923-933.

Cavalli-Sforza L and Edwards AW (1966) Estimating procedures for evolutionary branching processes. *Bulletin of the International Statistical Institute* **41**:803-808.

Cavalli-Sforza L and Edwards AW (1967) Phylogenetic analysis: models and estimation procedures. *Evolution* **21**:550-570.

Conway-Morris S (1997) Defusing the Cambrian 'explosion'? *Current Biology* **7**:R71-R74.

Conway-Morris S (1998a) Metazoan phylogenies: falling into place or falling to pieces? A palaeontological perspective. *Current Opinion in Genetics and Development* **8**:662-667.

Conway-Morris S (1998b) *The crucible of creation: the Burgess Shale and the rise of animals*. Oxford: Oxford University Press.

Cooper A and Fortey R (1998) Evolutionary explosions and the phylogenetic fuse. *Trends in Ecology and Evolution.* **13**:151-156.

Cooper A and Penny D (1997) Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* **275**:1109-1113.

Cox DR and Miller HD (1965) *The theory of stochastic processes.* Boca Raton: Chapman & Hall.

Crimes TP, Insole A, and Williams BPJ (1995) A rigid-bodied Ediacaran biota from Upper Cambrian strata in Co. Wexford, Eire. *Geological Journal* **30**:89-109.

Cutler DJ (2000a) Estimating divergence dimes in the presence of an overdispersed molecular clock. *Molecular Biology and Evolution* **17**:1647-1660.

Cutler DJ (2000b) The index of dispersion of molecular evolution: slow fluctuations. *Theoretical Population Biology* **57**:177-186.

Cutler DJ (2000c) Understanding the overdispersed molecular clock. *Genetics* **154**:1403-1417.

Darwin C (1859) *On the origin of species.* London: Murray.

Dempster AP, Schatzoff M, and Wermuth N (1977) A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association* **72**:77-106.

Doolittle RF and Blombäck B (1964) Amino-acid sequence investigations of fibrinopeptides from various mammals. *Nature* **202**:147-152.

Dopazo J (1994) Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. *Journal of Molecular Evolution* **38**:300-304.

Drake JW and Baltz RH (1976) The biochemistry of mutagenesis. *Annual Review of Biochemistry* **45**:11-37.

Easteal S (1999) Molecular evidence for the early divergence of placental mammals. *Bioessays* **21**:1052-1058.

Edwards AW (1970) Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society (B)* **32**:155-174.

Edwards AW (1992) *Likelihood.* 2nd ed. Baltimore and London: John Hopkins University Press.

Edwards AW and Cavalli-Sforza L (1963) The reconstruction of evolution. *Heredity* **18**:553.

Edwards AW and Cavalli-Sforza L (1964) Reconstruction of evolutionary trees. Phenetic and phylogenetic classification. *Systematic Association Publication* **6**:67-76.

Efron B (1979) Bootstrap methods: another look at the jacknife. *Annals of Statistics* **7**:1-26.

Efron B and Tibshirani R (1993) *An introduction to the bootstrap.* New York: Chapman & Hall.

Efron B and Tibshirani R (1998) The problem of regions. *Annals of Statistics* **26**:1687-1718.

Efron B, Halloran E, and Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences (USA)* **93**:13429-13434.

Eldredge N and Gould SJ (1972) Punctuated equilibria: an alternative to phyletic gradualism. In Schopf TMJ (ed) *Models in paleobiology.* San Francisco: Freeman.

Ewens WJ and Grant GR (2001) *Statistical methods in bioinformatics: an introduction.* New York: Springer-Verlag.

Fedonkin MA (1985) Precambrian metazoans: the problems of preservation, systematics and evolution. *Philosophical Transactions of the Royal Society London B Biological Sciences* **311**:27-44.

Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22**:240-249.

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**:401-410.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376.

Felsenstein J (1983) Statistical inference of phylogenies. *Journal of the Royal Statistical Society (A)* **146**:246-272.

Felsenstein J (1984) Distance methods for inferring phylogenies: a justification. *Evolution* **38**:16-24.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783-791.

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**:521-565.

Felsenstein J (1995) PHYLIP (phylogeny inference package). Distributed by the author, Department of Genetics, University of Washington, Seattle.

Felsenstein J (2001) Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution* **53**:447-455.

Felsenstein J and Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* **13**:93-104.

Felsenstein J and Kishino H (1993) Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology* **42**:193-200.

Feng DF, Cho G, and Doolittle RF (1997) Determining divergence times with a protein clock: update and reevaluation. *Proceedings of the National Academy of Sciences (USA)* **94**:13028-13033.

Fitch WM (1971) Rate of change of concomitantly variable codons. *Journal of Molecular Evolution* **1**:84-96.

Fitch WM (1976) Molecular evolutionary clocks. In Ayala FJ (ed) *Molecular evolution.* Sunderland, MA: Sinauer Associates.

Fitch WM and Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* **4**:579-593.

Force A, Lynch M, Pickett FB, Amores A, Yan Yl, and Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531-1545.

Fortey R (2001) The Cambrian explosion exploded? *Science* **293**:438-439.

Fortey RA, Briggs DE, and Wills MA (1996) The Cambrian evolutionary 'explosion': decoupling cladogenesis from morphological disparity. *Biological Journal of the Linnaean Society London* **57**:13-33.

Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution* **18**:866-873.

Galtier N and Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. *Proceedings of the National Academy of Sciences (USA)* **92**:11317-11321.

Galtier N and Gouy M (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution* **15**:871-879.

Galtier N, Tourasse N, and Gouy M (1999) A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**:220-221.

Gaut BS and Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular Biology and Evolution* **12**:152-162.

George E and McCulloch R (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**:881-889.

Geyer CJ (1999) Likelihood inference for spatial point processes. In Barndorff-Nielsen OE, Kendall WS, van Lieshout MNM (eds) *Stochastic geometry. Likelihood and computation*. Boca Raton: Chapman & Hall.

Gilks WR, Richardson S, and Spiegelhalter DJ (1996) *Markov chain Monte Carlo in practice.* Boca Raton: Chapman & Hall.

Gill PE, Murray W, and Wright MH (1981) *Practical optimization.* London: Academic Press.

Gillespie JH (1988) More on the overdispersed molecular clock. *Genetics* **118**:385-388.

Gillespie JH (1991) *The causes of molecular evolution.* Oxford: Oxford University Press.

Glaessner MF (1984) *The dawn of animal life: a biohistorical study.* Cambridge: Cambridge University Press.

Golden RM (1995) Making correct statistical inferences using a wrong probability model. *J. Math. Psych.* **39**:3-20.

Goldman N (1993) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**:182-198.

Goldman N (1994) Variance to mean ratio, R(t), for Poisson processes on phylogenetic trees. *Molecular Phylogenetics and Evolution* **3**:230-239.

Goldman N and Whelan S (2000) Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Molecular Biology and Evolution* **17**:975-978.

Goldman N and Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**:725-736.

Goldman N, Anderson JP, and Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. *Systematic Biology* **49**:652-670.

Goodman M, Koop BF, Czelusniak J, and Weiss ML (1984) The eta-globin gene. Its long evolutionary history in the beta-globin gene family of mammals. *Journal of Molecular Biology* **180**:803-823.

Gould SJ. 2001. *The lying stones of Marrakech.* London: Vintage.

Gu X (1998) Early metazoan divergence was about 830 million years ago. *Journal of Molecular Evolution* **47**:369-371.

Gu X, Fu YX, and Li WH (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution* **12**:546-557.

Haeckel E (1866) *Generelle Morphologie der Organismen.* Berlin: Georg Riemer.

Hall P and Wilson SR (1991) Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**:757-762.

Hasegawa M and Kishino H (1989) Confidence limits on the maximum likelihood estimation of the hominoid tree for mitochondrial DNA sequences. *Evolution* **43**:672-677.

Hasegawa M and Kishino H (1994) Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree. *Molecular Biology and Evolution* **11**:142-145.

Hasegawa M, Di Rienzo A, Kocher TD, and Wilson AC (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. *Journal of Molecular Evolution* **37**:347-354.

Hasegawa M, Kishino H, and Saitou N (1991) On the maximum likelihood method in molecular phylogenetics. *Journal of Molecular Evolution* **32**:443-445.

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

Hasegawa M, Kishino H, and Yano T (1987) Man's place in Hominoidea as inferred from molecular clocks of DNA. *Journal of Molecular Evolution* **26**:132-147.

Hasegawa M, Kishino H, and Yano T (1989) Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. *Journal of Human Evolution* **18**:476.

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

Haughton DMA (1988) On the choice of a model to fit data from an exponential family. *Annals of Statistics* **16**:342-355.

Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL, and Hedges SB (2001) Molecular evidence for the early colonization of land by fungi and plants. *Science* **293**:1129-1133.

Hennig W (1966) *Phylogenetic systematics.* Urbana: Illinois University Press.

Hillis DM and Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**:182-192.

Hillis DM, Moritz C, and Mable BK (1996) *Molecular systematics*. Swofford DL et al (eds).2nd edition. Sinauer, Sunderland, Mass.

Hoffman PF (1991) Did the breakout of Laurentia turn Gondwanaland inside-out? *Science* **252**:1409-1412.

Hoffman PF, Kaufman AJ, Halverson GP, and Schrag DP (1998) A neoproterozoic snowball earth. *Science* **281**:1342-1346.

Holmes SP (1999) Phylogenies: an overview. In Halloran ME, Geisser S (eds) *Statistics in genetics*. New York: Springer.

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, and Takahata N (1992) Man's place in Hominoidea revealed by mitochondrial DNA genealogy. *Journal of Molecular Evolution* **35**:32-43.

Huelsenbeck JP and Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics* **28**:437-466.

Huelsenbeck JP and Imennov NS (2002) Geographic origin of human mitochondrial DNA: accommodating phylogenetic uncertainty and model comparison. *Systematic Biology* **51**:155-165.

Huelsenbeck JP and Nielsen R (1999) Variation in the pattern of nucleotide substitution across sites. *Journal of Molecular Evolution* **48**:86-93.

Huelsenbeck JP and Rannala B (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**:227-232.

Huelsenbeck JP and Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. **17**:754-755.

Huelsenbeck JP Larget B and Swofford DL (2000a) A compound Poisson process for relaxing the molecular clock. *Genetics* **154**:1879-1892.

Huelsenbeck JP, Hillis DM, and Nielsen R (1996) A likelihood-ratio test of monophyly. *Systematic Biology* **45**:544-556.

Huelsenbeck JP, Rannala B, and Masly JP (2000b) Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288**:2349-2350.

Huelsenbeck JP, Ronquist F, Nielsen R, and Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**:2310-2314.

Hurst LD and Pal C (2001) Evidence for purifying selection acting on silent sites in BRCA1. *Trends in Genetics* **17**:62-65.

Hyde WT, Crowley TJ, Baum SK, and Peltier WR (2000) Neoproterozoic 'snowball Earth' simulations with a coupled climate/ice-sheet model. *Nature* **405**:425-429.

Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution* **40**:190-226.

Johnson NI and Kotz S (1973) *Discrete distributions.* Boston: Houghton Mifflin.

Jukes TH and Cantor CR (1969) Evolution of protein molecules. In Munro HN (ed) *Mammalian protein metabolism.* Academic Press, New-York.

Karlin S and Taylor HM (1981) *A secondary course in stochastic processes.* New York: Academic Press.

Kass RE and Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association.* **90**:773-795.

Kass RE and Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**:928-934.

Kelly C and Rice J (1996) Modeling nucleotide evolution: a heterogeneous rate analysis. *Mathematical Biosciences* **133**:85-109.

Kempthorne O and Folks L (1971) *Probability, statistics, and data analysis.* Ames: Iowa State University Press.

Kendall DG (1948) On the generalized birth-and-death process. *Annals of Mathematical Statistics* **19**:1-15.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature* **217**:624-626.

Kimura M (1969) The rate of molecular evolution considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences (USA)* **63**:1181-1188.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**:111-120.

Kimura M (1983) *The neutral theory of molecular evolution.* Cambridge: Cambridge University Press.

King JL and Jukes TH (1969) Non-Darwinian evolution. *Science* **164**:788-798.

Kirschvink JL, Ripperdan RL, and Evans DA (1997) Evidence for a large-scale reorganization of Early Cambrian continental masses by inertial interchange true polar wander. *Science* **277**:541-545.

Kishino H and Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution* **29**:170-179.

Kishino H and Hasegawa M (1990) Converting distance to time: application to human evolution. *Methods in Enzymology.* **183**:550-570.

Kishino H, Miyata T, and Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution* **30**:151-160.

Kishino H, Thorne JL, and Bruno WJ (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* **18**:352-361.

Knoll AH and Carroll SB (1999) Early animal evolution: emerging views from comparative biology and geology. *Science* **284**:2129-2137.

Kocher TD and Wilson AC (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and protein-coding regions. In Osawa S, Honio T (eds). Tokyo: Springer Verlag.

Kohne DE (1970) Evolution of higher-organism DNA. *Q.Rev.Biophys.* **3**:327-375.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, and Bhattacharya T (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**:1789-1796.

Kullback S and Leibler RA (1951) On information and sufficiency. *Annals of Mathematical Statistics* **22**:79-86.

Kumar S and Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature* **392**:917-920.

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences (USA)* **91**:1455-1459.

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

Langley CH and Fitch WM (1974) An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* **3**:161-177.

Larget B and Simon D (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* **16**:750-759.

Lee MS (1999) Molecular clock calibrations and metazoan divergence dates. *Journal of Molecular Evolution* **49**:385-391.

Lee MS (2001) Unalignable sequences and molecular evolution. *Trends in Ecology and Evolution* **16**:681-685.

Lehmann EL (1959) *Testing statistical hypotheses.* New York: Wiley.

Leitner T and Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proceedings of the National Academy of Sciences (USA)* **96**:10752-10757.

Lemieux C, Otis C, and Turmel M (2000) Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. *Nature* **403**:649-652.

Li P and Bousquet J (1992) Relative-rate test for nucleotide substitutions between two lineages. *Molecular Biology and Evolution* **9**:1185-1189.

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**:96-99.

Li WH (1997) *Molecular evolution.* Sunderland Massuchusetts: Sinauer.

Li WH and Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326**:93-96.

Li WH, Ellsworth DL, Krushkal J, Chang BH, and Hewett-Emmett D (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Molecular Phylogenetics and Evolution* **5**:182-187.

Li WH, Wu CI, and Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**:150-174.

Linhart H (1988) A test whether two AIC's differ significantly. *South African Statistical Journal* **22**:153-161.

Lynch M and Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**:459-473.

Madigan D and Raftery AE (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**:1535-1546.

Madigan D and York J (1995) Bayesian graphical models for discrete data. *International Statistical Review* **63**:215-232.

Marais G and Duret L (2001) Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. *Journal of Molecular Evolution* **52**:275-280.

Margoliash E (1963) Primary structure and evolution of cytochrome C. *Proceedings of the National Academy of Sciences (USA)* **50**:672-679.

Martin AP (2001) Molecular clocks. *Encyclopaedia of life sciences*. London: Nature Publishing Group.

Martin AP and Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences (USA)* **90**:4087-4091.

Martin AP, Naylor GJ, and Palumbi SR (1992) Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. *Nature* **357**:153-155.

McGuire G, Denham MC, and Balding DJ (2001) Models of sequence evolution for DNA sequences containing gaps. *Molecular Biology and Evolution* **18**:481-490.

McMenamin MAS (1989) The origins and radiation of the early metazoa. In Allen K, Briggs DE (eds) *Evolution and the fosil record*. London: Bellheaven Press.

Metropolis NA, Rosenbluth AW, Rosenbluth NM, Teller AH, and Teller E (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**:1087-1091.

Miyata T and Suga H (2001) Divergence pattern of animal gene families and relationship with the Cambrian explosion. *Bioessays* **23**:1018-1027.

Miyata T and 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. *Journal of Molecular Evolution* **16**:23-36.

Miyata T, Hayashida H, Kuma K, Mitsuyasu K, and Yasunaga T (1987) Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harbor Symposium on Quantitative Biology* **52**:863-867.

Mood AM, Graybill FA, and Boes DC (1974) *Introduction to the theory of statistics*. 3rd ed. Auckland: McGraw Hill.

Muse SV and Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**:715-724.

Muse SV and Weir BS (1992) Testing for equality of evolutionary rates. *Genetics* **132**:269-276.

Narbonne GM, Saylor BZ, and Grotzinger JP (1997) The youngest Ediacaran fossils from southern Africa. *Journal of Paleontology* **71**:953-967.

Nee S, May RM, and Harvey PH (1994) The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society London B Biological Sciences* **344**:305-311.

Nei M (1987) *Molecular evolutionary genetics*. New York: Columbia University Press.

Nei M and Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**:418-426.

Nei M and Kumar S (2000) *Molecular evolution and phylogenetics.* Oxford: Oxford University Press.

Neyman J (1971) Molecular studies of evolution: a source of novel statistical problems. In Gupta SS, Yackel J (eds) *Statistical decision theory and related topics*. New York: Academic Press.

Nielsen C (1995) *Animal evolution: interrelationships of the living phyla.* Oxford: Oxford University Press.

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

Nielsen R, Huelsenbeck JP (2002) Detecting positively selected amino-acid sites using posterior predictive p-values. In Altman RB et al (eds) *Pacific symposium on biocomputing 2002*. New Jersey: World Scientific.

Ohta T and Kimura M (1971) On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution* **1**:18-25.

Ota R, Waddell PJ, Hasegawa M, Shimodaira H, and Kishino H (2000) Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Molecular Biology and Evolution* **17**:798-803.

Pagel M (1997) Inferring the evolutionary process from phylogenies. *Zoologia Scripta* **26**:331-348.

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, and Stanley HE (1992) Long-range correlations in nucleotide sequences. *Nature* **356**:168-170.

Penny D (1982) Towards a basis for classification: the incompleteness of distance measures, incompatibility analysis and phenetic classification. *Journal of Theoretical Biology* **96**:129-142.

Penny D, McComish BJ, Charleston MA, and Hendy MD (2001) Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution* **53**:711-723.

Perlman MD and Wu L (1999) The Emperor's new tests. *Statistical Science* **14**:355-381.

Pitman EJG (1939) The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika* **30**:391-421.

Pollard SL and Holland PW (2000) Evidence for 14 homeobox gene clusters in human genome ancestry. *Current Biology* **10**:1059-1062.

Posada D and Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics.* **14**:817-818.

Posada D and Crandall KA (2001a) Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Molecular Biology and Evolution* **18**:897-906.

Posada D and Crandall KA (2001b) Selecting the best-fit model of nucleotide substitution. *Systematic Biology* **50**:580-601.

Raftery AE (1996) Hypothesis testing and model selection. In Gilks WR, Richardson S, Spiegelhalter DJ (eds) *Marko chain Monte Carlo in practice.* Boca Raton: Chapman & Hall.

Raftery AE and Lewis SM (1996) Implementing MCMC. In Gilks WR, Richardson S, Spiegelhalter DJ (eds) *Markov chain Monte Carlo in practice.* Boca Raton: Chapman & Hall.

Rambaut A and Bromham L (1998) Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution* **15**:442-448.

Rand DM (1994) Thermal habit, metabolic rate and the evolution of mitochondrial DNA. *Trends in Ecology and Evolution.* **9**:131.

Relethford JH (2001) Ancient DNA and the origin of modern humans. *Proceedings of the National Academy of Sciences (USA)* **98**:390-391.

Rogers JS (2001) Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Systematic Biology* **50**:713-722.

Rogers JS and Swofford DL (1999) Multiple local maxima for likelihoods of phylogenetic trees: a simulation study. *Molecular Biology and Evolution* **16**:1079-1085.

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

Runnegar B (1986) Molecular palaeontology. *Palaeontology.* **29**:1-24.

Runnegar B (2000) Loophole for snowball Earth. *Nature* **405**:403-404.

Rzhetsky A and Nei M (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution* **10**:1073-1095.

Saitou N (1988) Property and efficiency of the maximum likelihood method for molecular phylogeny. *Journal of Molecular Evolution* **27**:261-273.

Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **14**:1218-1232.

Sanderson MJ (1998) Estimating rate and time in molecular phylogenies: beyond the molecular clock? In Soltis PS, Soltis DE, Doyle JA (eds) *Plant molecular systematics,* 2nd ed. London: Chapman and Hall.

Sanderson, MJ, Wojciechowski MF, Hu J, Khan TS, and Brady SG (2000) Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Molecular Biology and Evolution*. **17**:782-797.

Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* **19**:101-109.

Sarich VM and Wilson AC (1967) Immunological time scale for hominid evolution. *Science* **158**:1200-1203.

Schoniger M and von Haeseler A (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution* **3**:240-247.

Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics* **6**:461-464.

Seilacher A (1985) Discussion of precambrian metazoans. *Philosophical Transactions of the Royal Society London B Biological Sciences* **311**:47-48.

Seilacher A, Bose PK, and Pfluger F (1998) Triploblastic animals more than 1 billion years ago: trace fossil evidence from india. *Science* **282**:80-83.

Self SG and Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**:605-610.

Shao J (1999) *Mathematical statistics.* New York: Springer.

Shimodaira H (1998) An application of multiple comparison techniques to model selection. *Annals of the Institute of Statistical Mathematics* **50**:1-13.

Shimodaira H (2001) Multiple comparisons of log-likelihoods and combining non-nested models with applications to phylogenetic tree selection. *Communications in Statistics, Part A-Theory and Methods*

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51**:492-508.

Shimodaira H and Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**:1114-1116.

Simpson GG (1944) *Tempo and mode in evolution.* New York: Columbia University Press.

Simpson GG (1964) Organisms and molecules in evolution. *Science* **146**:1535-1538.

Sitnikova T, Rzhetsky A, and Nei M (1995) Interior-branch and bootstrap tests of phylogenetic trees. *Molecular Biology and Evolution* **12**:319-333.

Siveter DJ, Williams M, and Waloszek D (2001) A phosphatocopid crustacean with appendages from the Lower Cambrian. *Science* **293**:479-481.

Slack JM, Holland PW, and Graham CF (1993) The zootype and the phylotypic stage. *Nature* **361**:490-492.

Sokal RR and Rolf FJ (1995) *Biometry : the principles and practice of statistics in biological research.* 3rd ed. New York: Freeman.

Sokal RR and Sneath PH (1963) *Principles of numerical taxonomy.* San Francisco and London: Freeman and company.

Springer MS (1995) Molecular clocks and the incompleteness of the fossil record. *Journal of Molecular Evolution* **41**:531-538.

Steel M (1994) The maximum-likelihood point for a phylogenetic tree is not unique. *Systematic Biology* **43**:560-564.

Strimmer K and Pybus OG (2001) Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution* **18**:2298-2305.

Sullivan J and 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? *Systematic Biology* **50**:723-729.

Swofford DL, Olsen GJ, Waddell PG, and Hillis DM (1996) Phylogenetic inference. In Hillis DM, Moritz C, Mable BK (eds) *Molecular systematics*. Sunderland, Massachusetts: Sinauer.

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, and Rogers JS (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology* **50**:525-539.

Tajima F (1993a) Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**:599-607.

Tajima F (1993b) Unbiased estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution* **10**:677-688.

Takahashi K and Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Molecular Biology and Evolution* **17**:1251-1258.

Takahata N (1987) On the overdispersed molecular clock. *Genetics* **116**:169-179.

Takahata N (1991a) Overdispersed molecular clock at the major histocompatibility complex loci. *Proceedings of the Royal Society London B Biological Sciences* **243**:13-18.

Takahata N (1991b) Statistical models of the overdispersed molecular clock. *Theoretical Population Biology* **39**:329-344.

Takezaki N, Rzhetsky A, and Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution* **12**:823-833.

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

Tateno Y, Takezaki N, and Nei M (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution* **11**:261-277.

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**:57-86.

Thompson EA (1975) *Human evolutionary trees.* Cambridge: Cambridge University Press.

Thompson JD, Higgins DG, and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**:4673-4680.

Thorne JL, Kishino H, and Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**:1647-1657.

Urrutia AO and Hurst LD (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159**:1191-1199.

Uyenoyama MK (1995) A generalized least-squares estimate for the origin of sporophytic self-incompatibility. *Genetics* **139**:975-992.

Uzzell T and Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. *Science* **172**:1089-1096.

Valentine JW (1994) Late Precambrian bilaterians: grades and clades. *Proceedings of the National Academy of Sciences (USA)* **91**:6751-6757.

Valentine JW, Erwin DH, and Jablonski D (1996) Developmental evolution of metazoan bodyplans: the fossil evidence. *Developmental Biology* **173**:373-381.

Valentine JW, Jablonski D, and Erwin DH (1999) Fossils, molecules and embryos: new perspectives on the Cambrian explosion. *Development* **126**:851-859.

Vermeij GJ (1996) Animal origins. *Science* **274**:525-526.

Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica.* **57**:307-333.

Wakeley J (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *Journal of Molecular Evolution* **37**:613-623.

Wald A (1949) Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **20**:595-601.

Wallace A (1997) *The origin of animal body plans: a study in evolutionary developmental biology.* Cambridge: Cambridge University Press.

Wang DY, Kumar S, and Hedges SB (1999) Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proceedings of the Royal Society London B Biological Sciences* **266**:163-171.

Waterman MS (1995) *Introduction to computational biology.* London: Chapman & Hall.

Westfall PH and Young SS (1993) *Resampling-based multiple testing: examples and methods for p-value adjustments.* New York: John Wiley & Sons.

Whelan S and Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Molecular Biology and Evolution* **16**:1292-1299.

Whelan S, Lio P, and Goldman N (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics* **17**:262-272.

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**:60-62.

Woodroofe M (1982) On model selection and the ARC sine laws. *Annals of Statistics* **10**:1182-1194.

Wray GA, Levinton JS, and Shapiro LH (1996) Molecular evidence for deep Precambrian divergences. *Science* **274**:568-573.

Wu CI and Li WH (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences (USA)* **82**:1741-1745.

Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10**:1396-1401.

Yang Z (1994a) Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**:105-111.

Yang Z (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**:306-314.

Yang Z (1995) A space-time process model for the evolution of DNA sequences. *Genetics* **139**:993-1005.

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

Yang Z (1996b) Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution* **42**:587-596.

Yang Z (1997a) How often do wrong models produce better phylogenies? *Molecular Biology and Evolution* **14**:105-108.

Yang Z (1997b) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**:555-556.

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**:568-573.

Yang Z (2000a) Complexity of the simplest phylogenetic estimation problem. *Proceedings of the Royal Society London B Biological Sciences* **267**:109-116.

Yang Z (2000b) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of Molecular Evolution* **51**:423-432.

Yang Z (2001) Adaptive molecular evolution. In Balding DJ, Bishop M, Cannings C (eds) *Handbook of statistical genetics*. London: Wiley.

Yang Z and Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**:409-418.

Yang Z and Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**:32-43.

Yang Z and Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. *Molecular Biology and Evolution* **14**:717-724.

Yang Z and Roberts D (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution* **12**:451-458.

Yang Z and Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution* **19**:49-57.

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

Yang Z, Goldman N, and Friday A (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Systematic Biology* **44**:384-399.

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

Yoder AD and Yang Z (2000) Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* **17**:1081-1090.

Zhang J (1999) Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Molecular Biology and Evolution* **16**:868-875.

Zharkikh A and Li WH (1992a) Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution* **9**:1119-1147.

Zharkikh A and Li WH (1992b) Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. Four taxa without a molecular clock. *Journal of Molecular Evolution* **35**:356-366.

Zuckerkandl E and Pauling L (1962) Molecular disease, evolution, and genetic heterogeneity. In Kasha M, Pullman B (eds) *Horizons in biochemistry*. New York: Academic Press.

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

Appendices

Appendix 1 – PHYBAYES: a program for phylogenetic analyses in a Bayes framework

General description: PHYBAYES is a multiplatform program that implements Bayesian methods in phylogenetics for analysing nucleotide sequences. This program has two singular features: (i) it allows the estimation of divergence times and rates of evolution when the molecular clock assumption is relaxed; (ii) it computes the probability of the data under a given model, from which the Bayes factor can be obtained to construct confidence sets of trees or to test hypotheses.

PHYBAYES is a program written in ANSI C which implements a Bayesian approach for nucleotide-based phylogenetic analyses. While the main purpose is to estimate divergence times when the molecular clock assumption does not hold, PHYBAYES can also compare models in a Bayesian framework and test phylogenetic hypotheses.

The molecular clock assumption is relaxed by choosing a model of rate change. I have shown this to lead to estimates of divergence times different from and better than those obtained under the clock (Chapter II). With PHYBAYES, different prior distributions can be placed on divergence times: a uniform distribution, a beta distribution, or a birth-death process with species sampling (Yang and Rannala, 1997). The parameters of these distributions are integrated out over uniform distributions, whose respective ranges are defined by the user in the control file. Likewise, several models of rate change are available. The rate of a branch is centred around the rate of its ancestral branch and its distribution can be lognormal (Thorne et al., 1998), "stationarised" lognormal (Kishino et al., 2001), gamma, exponential, or can follow the Ornstein-Uhlenbeck process. Hyperparameters can be integrated out (gamma prior distributions vague in the region where the model does not behave like the clock), or estimated using an empirical Bayes

approach. Marginal posterior distributions are approximated by Markov chain Monte Carlo (MCMC), using the Metropolis-Hastings algorithm (e.g. Gilks et al., 1996).

Like other implementations, PHYBAYES estimates the posterior probability of the topologies by searching the tree space (Yang and Rannala, 1997; Larget and Simon, 1999; Huelsenbeck and Ronquist, 2001) and each tree sampled along the Markov chain is stored with its branch lengths in a file. A natural extension of the Bayesian methodology is to compare possibly non-nested models, such as different models of rate change, or tree topologies. PHYBAYES uses the Bayes factor for this purpose, i.e. the ratio of the probability of the data under each model. In particular, one can test whether a specific tree such as the maximum likelihood tree is the correct tree, or construct a confidence set of trees (Chapter IV). The uncertainty about the parameters of the model (parameters of the nucleotide substitution model, hyperparameters for the models of speciation and the models of rate change) is taken into account by integrating them out over uninformative (uniform) prior distributions.

The sequence file follows the formats accepted by PHYLIP (Felsenstein, 1995) and PAML (Yang, 1997b). It is possible to set up partitions for different genes or different codon positions. Options of the program are set in a control file and are grouped into subsections delimited by headers. This file also contains the names of the input and output files, the random number generator seed and the substitution model. In particular, five Markov models of nucleotide substitution are implemented: JC69, K80, F81, F84 and HKY85 (see Yang, 1997b). Their respective parameters can be fixed or integrated out, assuming they follow uninformative prior distributions. Among-site rate variation can be accommodated by a discrete gamma distribution (Yang, 1994b) for which the number of categories is set by the user. The run settings of the MCMC, such as the burn-

in period, the sample size (total number of samples kept for inference) and the thinning of the chain (sampling once every *k* accepted states), are also defined in the control file.

Depending on the parameters on which the MCMC is run, PHYBAYES outputs the corresponding posterior distributions in separate files. As an example, I have estimated the divergence times of six mammals. The orang-utan split 13 million years ago (MYA) is used to calibrate the tree. Figure A1.1 presents the posterior distribution of the divergence times scaled and plotted for the most probable topology, which PHYBAYES indicated to be ((((human, chimpanzee),orang-utan),(mouse, rat)),platypus). The analysis under the clock dates the mouse and rat split to ca. 50 MYA and the primates and rodents divergence to ca. 115 MYA (Figure A1.1). Note that the molecular clock is rejected by a LRT. Under an exponential model of rate change, mouse and rat split ca. 30 MYA, with a 95% credible set (95cs) of 16 – 72 MYA, while primates and rodents diverged around 65 MYA (95cs: 47 – 157 MYA). This is closer to common wisdom or molecular studies based on protein coding genes where the clock holds (Tavaré, 1986).

When the chain is run on the tree space, the sampled trees with their branch lengths are collected in a separate file (nexus format), that can be used as an input file for studies evaluating the impact of uncertainties such as tree topology or the parameters of the substitution model on the branch lengths. The probability of the data under the model, used to compute the Bayes factor, and summary statistics, are stored.

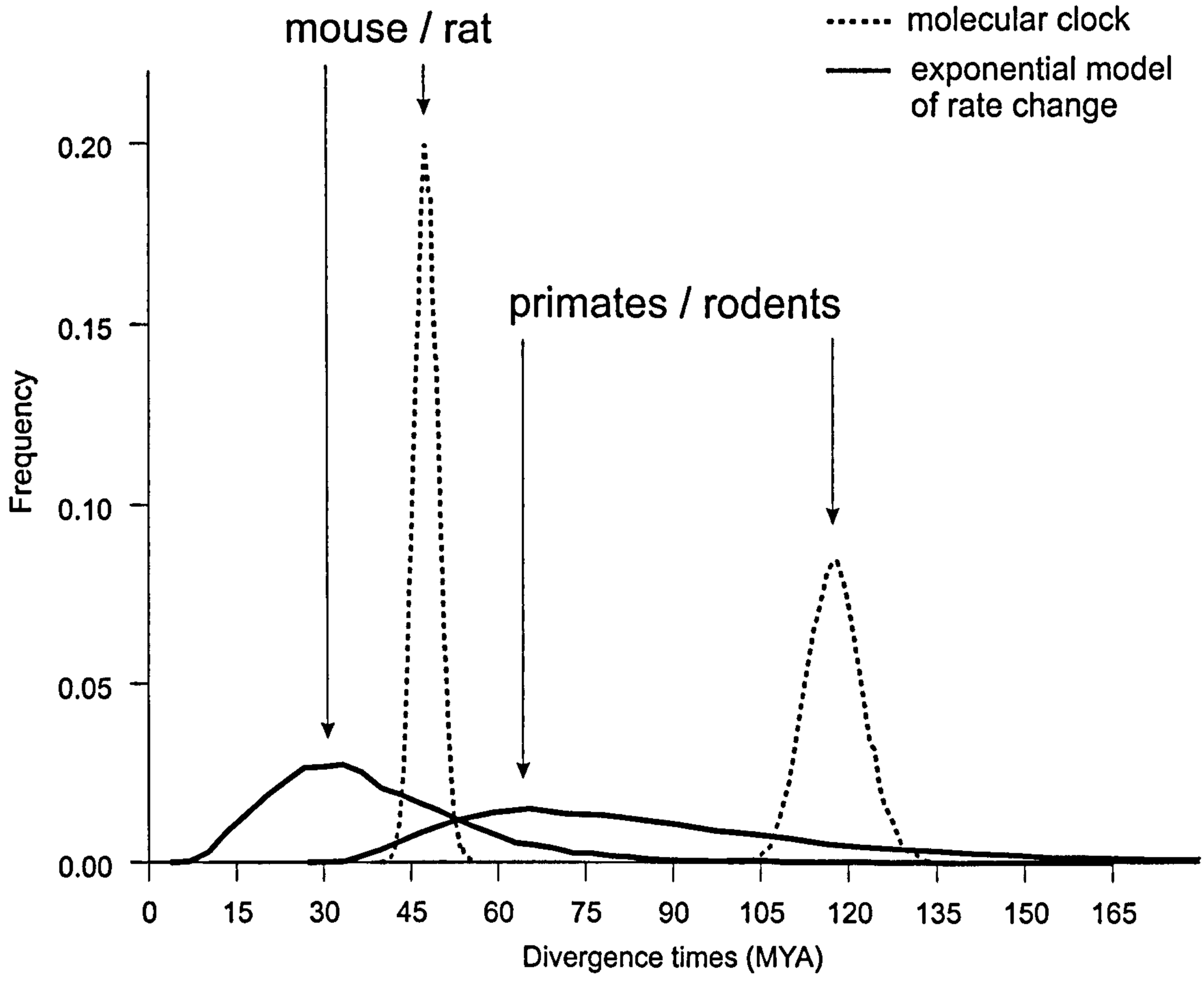


Figure A1.1. Posterior distributions of the divergence times under the most probable tree, as calculated by PHYBAYES, under two model of rate change: clock (broken lines) and exponential (solid lines). The dataset consists of the nucleotide sequences of the complete mitochondrial genome (10,806 bp) of six mammalian species: H = *Homo sapiens* (human), C = *Pan troglodytes* (common chimpanzee), O = *Pongo pygmaeus* (orang-utan), M = *Mus musculus* (mouse), R = *Rattus rattus* (rabbit) and P = *Platypus compertus* (platypus). The HKY85 + Γ substitution model is assumed. The calibration point is the orang-utan divergence, 13 MYA. A birth-death prior was assumed for the speciation process. After a burn-in of 10^5, the MCMC was sampled every 500 steps until 10^4 states were collected.

Although running time can become long, PHYBAYES has been tested on a fairly large data set (Chapter II) consisting of 40 species with 1,032 site patterns. Depending on the models, runs take two to three weeks on a Pentium III (1 GHz) with a burn-in period of 10^4 steps and sampling every 100 accepted steps until 10^4 states are collected. Computational time increases with the number of taxa, the number of site patterns, and the complexity of the model. Starting and tuning parameters should be optimised by running preliminary short chains. Convergence must be checked systematically. The program is distributed freely (http://abacus.gene.ucl.ac.uk/stephane/) and comes with a manual and example files.

Appendix 2 – Genes used for the large scale analysis

Mitochondrial genes

cox1, cox2, cox3, cytB, nad1, nad2, nad3, nad4, nad4L, nad5, nad6: *Alligator mississippiensis* (NC_001922); *Anopheles quadrimaculatus* (NC_000875); *Arabidopsis thaliana* (NC_001284); *Arbaxia lixula* (MIALDNASQ); *Asterina pectinifera* (NC_001627); *Balaenoptera musculus* (NC_001601); *Ceratotherium simum* (NC_001808); *Crassostrea gigas* (NC_001276); *Drosophila melanogaster* (NC_001709); *Echinococcus granulosus* (AF297617); *Eptatretus burgeri* (NC_002807);*Equus asinus* (NC_001788); *Florometra serratissima* (NC_001878); *Gallus gallus* (NC_001323); *Hymenolepis diminuta* (NC_002767); *Iguana iguana* (NC_002793); *Latimeria chalumnae* (NC_001804); *Limulus polyphemus* (AF216203); *Lithobius forficatus* (NC_002629); *Loligo bleekeri* (NC_002507); *Lumbricus terrestris* (LTU24570); *Metridium senile* (NC_000933); *Myxine glutinosa* (MGL404477);*Ornithorhynchus anatinus* (NC_000891); *Oryctolagus cuniculus* (NC_001913); *Paracentrotus lividus* (PALMTCG); *Platynereis dumerilii* (NC_000931); *Rhipicephalus sanguineus* (NC_002074); *Salmo salar* (AF133701); *Sardinops melanostictus* (NC_002616); *Strongylocentrotus purpuratus* (NC_001453); *Terebratulina retusa* (TRE245743); *Tetrodontophora bielanensis* (AF272824); *Tinamus major* (NC_002781); *Trichinella spiralis* (NC_002681).

Nuclear genes

18S: the alignment used is from Bromham et al. (1998).

actin: *Sus scrofa* (U16368); *Gallus gallus* (V01507); *Homo sapiens* (BC009978); *Salmo trutta* (AF267496); *Anolis carolinensis* (AF199487); *Equus caballus* (AF035774); *Marsupenaeus japonicus* (AB055975); *Anopheles gambiae* (U02933); *Limulus*

polyphemus (Z38130); *Arabidopsis thaliana* (U39449); *Podocoryne carnea* (X69059); *Hydra vulgaris* (M32364); *Penaeus monodon* (AF100987); *Aedes aegypti* (U20287).

α – tubulin: *Urechis caupo* (U30467); *Paracentrotus lividus* (X53618); *Hirudo medicinalis* (U67676); *Danio rerio* (AF029250); *Xenopus laevis* (X07046); *Oncorhynchus nerka* (AY026060); *Mus musculus* (M28727); *Meriones unguiculatus* (AF052694); *Ovis aries* (AF251146); *Chionodraco rastrospinosus* (AF263277); *Macaca fascicularis* (X04757); *Gecarcinus lateralis* (U92646); *Drosophila melanogaster* (M14644); *Caenorhabditis elegans* (AF003387); *Haemonchus contortus* (L02108); *Chironomus tentans* (AF272829); *Artemia franciscana* (AF078670); *Trypanosoma cruzi* (AF091836); *Toxoplasma gondii* (M20024); *Plasmodium falciparum* (X15979); *Arabidopsis thaliana* (M17189).

β – tubulin: *Strongylocentrotrus purpuratus* (X07502); *Paracentrotus lividus* (X15389); *Rattus norvegicus* (AB011679); *Cricetulus griseus* (AF120325); *Macaca mulatta* (AF147880); *Homo sapiens* (AF141349); *Gallus gallus* (V00389); *Chionodraco rastrospinosus* (AF255955); *Notothenia coriiceps* (AF255555); *Drosophila erecta* (M16922); *Heliothis virescens* (U75868); *Octopus dofleini* (L10111); *Bombyx mori* (AB011069); *Halocynthia roretzi* (D89794); *Arabidopsis thaliana* (M20405); *Cylicocyclus nassatus* (AF283767); *Cyathostomum coronatum* (AF283764); *Onchocerca volvulus* (AF019886).

calreticulin: *Strongylocentrotus purpuratus* (AF177914); *Onchocerca volvulus* (M20565); *Danio rerio* (AF195882); *Arabidopsis thaliana* (AY045656); *Drosophila melanogaster* (AB000718); *Amblyomma americanum* (U07708); *Rattus norvegicus* (D78308); *Mus musculus* (BC003453); *Oryctolagus cuniculus* (J05138); *Homo sapiens* (BC007911); *Bos Taurus* (L13462); *Rana rugosa* (D78589).

catalase: *Homo sapiens* (AY028632); *Canis familiaris* (AB038231); *Mus musculus* (AY040626); *Rana rugosa* (AB031872); *Danio rerio* (AF170069); *Drosophila melanogaster* (U00145); *Caenorhabditis elegans* (U55384); *Plexaura homomalla* (AF003692); *Strongylocentrotus purpuratus* (AF035380); *Lytechinus variegates* (AF035381); *Arabidopsis thaliana* (U43147).

elongation factor 1: *Oryctolagus cuniculus* (AF035178); *Salmo salar* (AF321836); *Cricetulus longicaudatus* (D00522); *Mus musculus* (BC004067); *Bos taurus* (AB060107); *Homo sapiens* (BC010735); *Gallus gallus* (U46663); *Xenopus laevis* (AB040437); *Seriola quinqueradiata* (AB032900); *Artemia sp.* (M28020); *Hydra vulgaris* (Z68181); *Arabidopsis thaliana* (AF360167); *Dreissena polymorpha* (AJ250733).

histone H1: *Xenopus laevis* (M36655); *Mus musculus* (AF034610); *Arabidopsis thaliana* (AY045797); *Rattus norvegicus* (M28409); *Gallus gallus* (M17020);*Bufo bufo* (AF255740); *Homo sapiens* (D64142); *Strongylocentrotus purpuratus* (M16033); *Tigriopus californicus* (M84797); *Rhynchosciara americana* (AF378198); *Drosophila virilis* (L76558); *Chironomus thummi* (L28731); *Chaetopterus variopedatus* (U96764).

heat shock protein 70: *Rattus norvegicus* (L16764); *Bos taurus* (U09861); *Homo sapiens* (X51758); *Danio rerio* (AF210640); *Gallus gallus* (J02579); *Crassostrea gigas* (AF144646); *Paracentrotus lividus* (X61379); *Biomphalaria glabrata* (AF025477); *Stylophora pistillata* (AF152004); *Botryllus schlosseri* (U51901); *Takifugu rubripes* (Y08577); *Arabidopsis thaliana* (AF217458).

protein kinase C: *Aplysia californica* (M94884); *Caenorhabditis elegans* (U00181); *Drosophila melanogaster* (J04848); *Homo sapiens* (AF345987); *Oryctolagus cuniculus* (M19338); *Bos taurus* (M13973); *Lytechinus pictus* (U02967); *Danio rerio* (AF390109); *Xenopus laevis* (U12588); *Blumeria graminis* (AF283107); *Mus musculus* (D11091).

troponin C: Patinopecten yessoensis (AB034963); Akazara scallop (D85883); Todarodes pacificus (AB049962); Oryctolagus cuniculus (J03462); Mus musculus (M57590); Danio rerio (AF180890); Xenopus laevis (AB003080); Gallus gallus (D13037); Homo sapiens (M37984); Perinereis vancaurica (AB052102); Drosophila silvestris (AF047329); Lytechinus pictus (J04068); Schizosaccharomyces pombe (AL035075).

Appendix 3 – Significance and hypothesis tests in a frequentist framework

In Chapter IV, I have presented four new tests. The emphasis of the thesis is on Bayes methods, and two Bayes tests were introduced to test molecular phylogenies either in a significance test approach (TBS) or in a hypothesis test approach (TBH). As pointed out by Nick Goldman, the frequentist equivalents of these tests, TFS and TFH, have an error in the test statistic used during the bootstrapping stage of the algorithm. In this appendix, I present the correction suggested by Nick Goldman so that the resampling matches more closely the first guideline of Hall and Wilson (1991).

The first guideline stresses that "*care should be taken to ensure that (...) resampling is done in a way that reflects* H_0" (Hall and Wilson, 1991, p.757). Let us consider testing H_0: $\theta = \theta_0$ against the two-sided alternative H_1: $\theta \neq \theta_0$. Let $\hat{\theta}$ be an estimator of the unknown quantity θ, and $\hat{\theta}^*$ be the value of $\hat{\theta}$ computed from the bootstrapped samples. Testing H_0 against H_1 is based on the unknown distribution of $\hat{\theta} - \theta_0$ under H_0, estimated by the distribution of $\hat{\theta}^* - \hat{\theta}$ (see I.3.c).

In Chapter IV, θ_0, $\hat{\theta}$ and $\hat{\theta}^*$ were respectively identified with c_λ, $\ell(T_i)$ and $\ell(T_i^*)$, where the notations are those of Fig IV.1. However, the null hypothesis should be defined prior to observing the data, and should therefore be independent of any prior analysis. This is not the case if θ_0 is identified with c_λ when defining the test statistics t_i's.

The correct test statistic $\hat{\theta}$, defined a priori, is actually $\ell(T_i) - c_\lambda$ (for simplicity's sake, I do not consider pivotal quantities for the moment), which leads to the identification of θ_0 with 0. This should be taken into account when computing the test statistics from the bootstrapped data, by calculating $\{\ell(T_i^*) - c_\lambda^*\} - \{\ell(T_i) - c_\lambda\}$, instead of $\{\ell(T_i^*) - \ell(T_i)\}$. The value of c_λ^* from the bootstrapped samples is $\sum_i^k \ell(T_i^*) / k$ for significance tests (TFS) and $\max_i \ell(T_i^*)$ for hypothesis tests (TFH).

As in Chapter IV, pivotal quantities are considered for the various test statistics. The tests described in Chapter IV will be referred to the "old" implementations hereafter; the procedure described in this appendix will be noted the "new" implementation.

When these changes are applied to the two small data sets of Chapter IV, the results obtained are somewhat different than the previous ones. Table A3.1 presents the results for the HIV-1 data set, and shows that, while the new hypothesis test gives the same confidence set of trees as the old implementation, this is not the case for the significance tests. Both T_2 and T_3 are excluded from the confidence set of trees at the lfd by the new implementation, irrespective of the assumed substitution model. Results of the new implementation for the larger mammalian mitochondrial data set are closer to the old procedure when the three codon positions are considered. However, when only the first two codon positions are analysed, the new implementation gives again smaller confidence sets, with respect to both the old implementation and existing selection procedures (BP and SH here). It is noticeable that the *p*-values obtained from the new implementations are more extreme than any other one, being either 0 or 1 in most of the cases.

In order to better understand some of the properties of the implemented test statistics, "old" and "new", I have performed some simulations. I concentrate essentially on the power of the different tests, as it is the quantity practitioners are usually interested in. I have used the codon positions one and two of the plastid gene *psbB* analysed by Sanderson et al. (2000). This is a highly conserved chloroplast photosystem gene, for which nineteen species have been sampled across the seed plants (see Sanderson et al. (2000) for background information) and 1,020 nucleotides are here analysed. In order to

Table A3.1. Comparison between the results of the old (as in Chapter IV) and new (as in this appendix) significance and hypothesis tests under the frequentist approach for the HIV-1 *gag* and *pol* data set.

Tree	$\Delta\ell$	BP	TFH-old	TFH-new	SH	TFS-old	TFS-new
HKY85+Γ							
T_{ML}	0.00	0.792	–	–	–	–	–
T_2	6.03	0.081	0.004	0.000	**0.160**	**0.166**	0.000
T_3	5.52	0.127	0.007	0.000	**0.190**	**0.225**	0.000
JC69							
T_{ML}	0.00	0.961	–	–	–	–	–
T_2	29.48	0.021	0.000	0.000	**0.035**	0.000	0.000
T_3	30.54	0.017	0.000	0.000	**0.027**	0.000	0.000

NOTE. –The topologies are those of Table IV.1. BP is the bootstrap proportion estimated by RELL. Non-significant results ($\alpha \geq 1\%$) are in bold.

Table A3.2a. Comparison between the results of the old (as in Chapter IV) and new (as in this appendix) significance and hypothesis tests under the frequentist approach for the mammalian mitochondrial data set when the three codon positions are considered.

Tree	$\Delta\ell$	BP	TFH-old	TFH-new	SH	TFS-old	TFS-new
T_{ML}	0.00	0.502	–	–	–	–	–
T_2	0.32	0.488	**0.359**	**0.026**	**0.844**	**1.000**	**1.000**
T_3	14.11	0.003	0.000	0.000	**0.344**	**1.000**	**1.000**
T_4	24.96	0.006	0.000	0.000	**0.093**	**0.999**	**1.000**
T_5	32.65	0.000	0.000	0.000	**0.019**	**0.991**	**1.000**
T_6	34.44	0.000	0.000	0.000	**0.020**	**0.771**	**1.000**
T_7	39.09	0.000	0.000	0.000	0.007	0.000	0.000
T_8	42.02	0.000	0.000	0.000	0.003	0.000	0.000
T_9	42.81	0.000	0.000	0.000	0.002	0.000	0.000
T_{10}	45.31	0.000	0.000	0.000	0.001	0.000	0.000
T_{11}	46.26	0.000	0.000	0.000	0.000	0.000	0.000
T_{12}	46.27	0.000	0.000	0.000	0.000	0.000	0.000
T_{13}	49.20	0.000	0.000	0.000	0.001	0.000	0.000
T_{14}	51.14	0.000	0.000	0.000	0.000	0.000	0.000
T_{15}	56.38	0.000	0.000	0.000	0.000	0.000	0.000

NOTE. –The topologies are those of Table IV.2. BP is the bootstrap proportion estimated by RELL. Non-significant results ($\alpha \geq 1\%$) are in bold.

Table A3.2b. Comparison between the results of the old (as in Chapter IV) and new (as in this appendix) significance and hypothesis tests under the frequentist approach for the mammalian mitochondrial data set when only codon positions one and two are considered.

Tree	$\Delta\ell$	BP	TFH-old	TFH-new	SH	TFS-old	TFS-new
T_{ML}	0.00	0.664	–	–	–	–	–
T_2	7.21	0.086	0.003	0.000	**0.556**	**1.000**	**1.000**
T_3	5.32	0.188	**0.020**	0.000	**0.652**	**1.000**	**1.000**
T_4	19.31	0.001	0.000	0.000	**0.094**	**0.728**	**1.000**
T_5	16.05	0.010	0.000	0.000	**0.172**	**0.963**	**1.000**
T_6	22.63	0.003	0.000	0.000	**0.052**	**0.215**	0.000
T_7	30.36	0.000	0.000	0.000	0.004	0.000	0.000
T_8	29.30	0.000	0.000	0.000	0.007	0.000	0.000
T_9	31.71	0.000	0.000	0.000	0.002	0.000	0.000
T_{10}	17.81	0.045	0.000	0.000	**0.133**	**0.878**	**1.000**
T_{11}	23.03	0.000	0.000	0.000	**0.041**	**0.173**	0.000
T_{12}	23.08	0.000	0.000	0.000	**0.041**	**0.166**	0.000
T_{13}	24.20	0.001	0.000	0.000	**0.034**	**0.083**	0.000
T_{14}	23.20	0.000	0.000	0.000	**0.036**	**0.155**	0.000
T_{15}	34.91	0.000	0.000	0.000	0.001	0.000	0.000

NOTE. –The topologies are those of Table IV.2. BP is the bootstrap proportion estimated by RELL. Non-significant results ($\alpha \geq 1\%$) are in bold.

encompass some complexity in the model, the data were analysed under REV + Γ. The parameter estimates were then used to simulate 500 data sets under REV + Γ with `evolver` (Yang, 1997b) for nineteen species of 1,020 nucleotides. The topology used to simulate the data was that estimated by Sanderson et al. (2000) for this partition of the *psbB* gene.

The topologies included in further analyses were selected from the nine trees with the highest posterior probability (with $p(T \mid X) > .01$) when HKY85 + Γ is assumed, plus the two trees selected under JC69, and a twelfth topology chosen a priori. The programme `bambe` (Larget and Simon, 1999) was used to perform the Bayes analyses.

To compare topologies and order them on a natural scale, we need to define a distance. As in a simulation study we know the generating model, the Kullback-Leibler distance (Kullback and Leibler, 1951) appears as a good candidate. This distance $d(f, g)$ between the generating model f and the approximating model g is defined as:

$$d(f,g) = \int f(x \mid \theta) \log \frac{f(x \mid \theta)}{g(x \mid \theta)} dx \qquad \text{(A3.1)}$$

which is the average of the logarithmic difference between the generating and the approximating model, with respect to the generating model:

$$d(f,g) = E_f[\log f(x \mid \theta)] - E_f[\log g(x \mid \theta)] \qquad \text{(A3.2)}$$

The estimated distance between f and g used hereafter is defined as:

$$D_{KL} = E\left[\log f(x \mid \hat{\theta})\right] - E\left[\log g(x \mid \hat{\theta})\right] \qquad \text{(A3.3)}$$

where the expectation is taken with respect to the 500 replicates, $f(\cdot)$ being the likelihood function of the topology used for simulating the data, and $g(\cdot)$ the likelihood function of any of the eleven other topologies selected above. The behaviour of the p-values and the power of the different tests are presented as functions of this distance. Power at level α was estimated from the simulations as the proportion of p-values less than α. I used

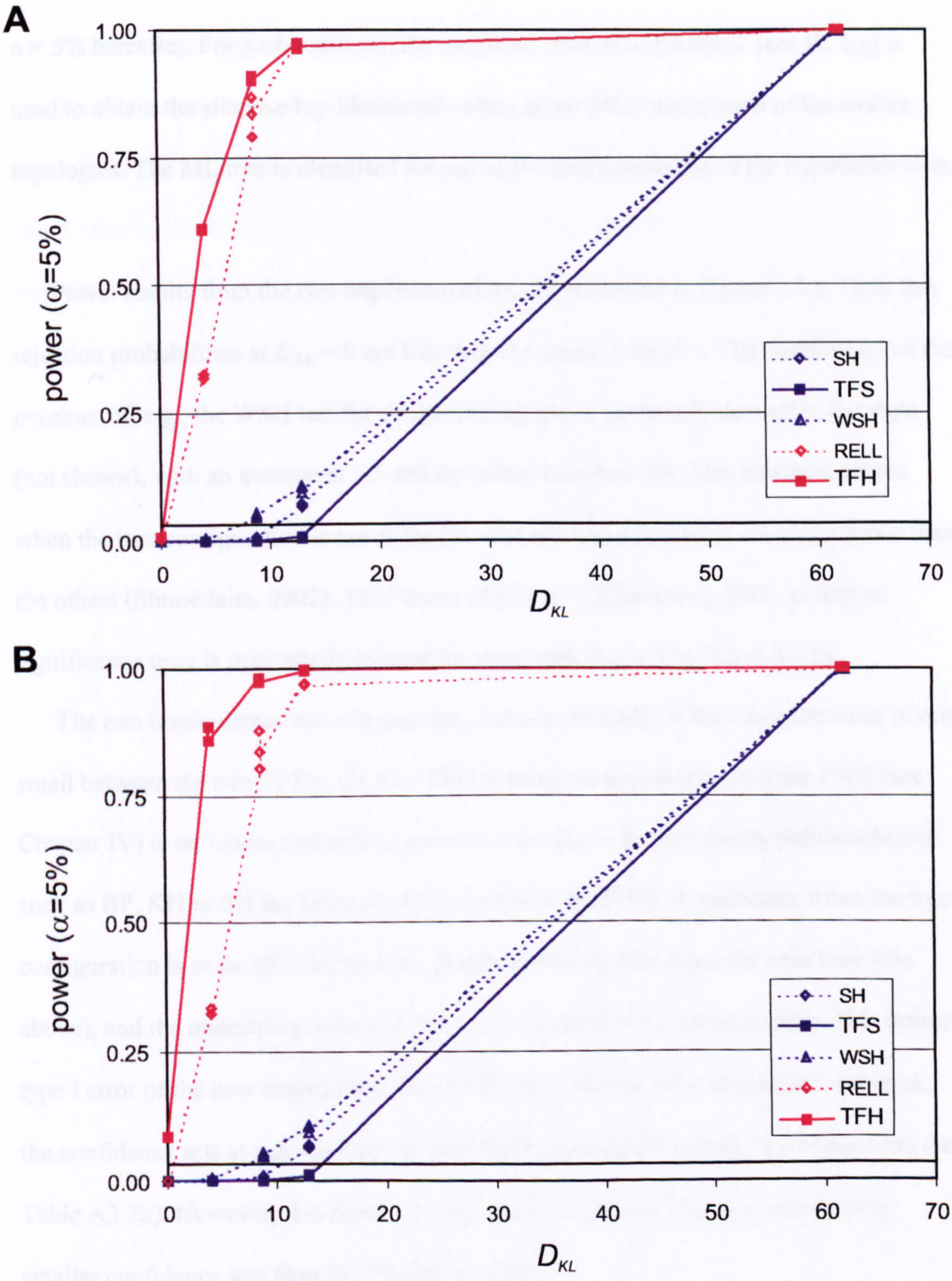


Fig. A3.1. Power curves of the existing tests (broken lines) and of the new tests (continuous lines) under their two implementation: **A**-old implementation of the test statistics; **B**-new implementation of the test statistics. See text for simulation details and the distance measure used between trees.

$\alpha = 5\%$ hereafter. For each replicate, the modified version of `baseml` (see IV.1.c) is used to obtain the sitewise log-likelihood values at the MLE under each of the twelve topologies. The ML tree is identified for use as the null hypothesis of the hypothesis tests.

Power results from the two implementations are presented in Figure A3.1. Note that rejection probabilities at $D_{KL} = 0$ are less than the nominal level α. The distribution of the *p*-values of, e.g., the WSH test for the generating tree is extremely skewed to the right (not shown), with an average at .95 and no values less than .48. This situation occurs when the true configuration is not at the lfd, that is when a few trees are much better than the others (Shimodaira, 2002). The "waste of power" (Shimodaira, 2002, p.500) of significance tests is particularly evident for trees with $D_{KL} < 13$ (Figure A3.1).

The two implementations, old and new, behave similarly. While the difference is very small between the two TFS's, the new TFH appears more powerful, but the FWE (see Chapter IV) is no longer controlled (power > α for $D_{KL} = 0$). Previously published tests such as BP, KH or SH are known to lack control of the FWE, in particular when the true configuration is at the lfd (Shimodaira, 2002). However, this is not the case here (see above), and the underlying reason of this lack of control is not entirely clear. This inflated type I error of the new implementation of FHT may explain why, in real data analyses, the confidence sets at a given level are smaller than those of existing hypothesis tests (see Table A3.2b). However, this does not explain why the new FTS has systematically smaller confidence sets than SH (Tables 1,2a,2b).

To conclude, both the correct implementation of the frequentist tests presented in Chapter IV and the corrected version of these tests have approximately the same power functions. Hypothesis tests are everywhere more powerful than significance tests, which

can exhibit a large type II error. However, in real data analysis, the confidence sets obtained from the new implementation tend to be smaller than either existing selection procedures (BP, SH) or than the old implementation. It is likely that the new tests are more sensitive to misspecification of the substitution model than existing selection procedures.