HAL open science

The Bayesian Approach to Molecular Phylogeny

Nicolas Lartillot

To cite this version:

Nicolas Lartillot. The Bayesian Approach to Molecular Phylogeny. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book, pp.1.4:1–1.4:17, 2020. hal-02535330

HAL Id: hal-02535330

https://hal.science/hal-02535330

Submitted on 10 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Chapter 1.4 The Bayesian Approach to Molecular Phylogeny

Nicolas Lartillot
Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive,
43 Bld du 11 Novembre 1918, 69622 Villeurbanne cedex, France.
nicolas.lartillot@univ-lyon1.fr
https://orcid.org/0000-0002-9973-7760

Abstract

Bayesian inference is now routinely used in phylogenomics and, more generally, in macroevolutionary studies. Beyond the philosophical debates it has raised concerning the choice of the prior and the meaning of posterior probabilities, Bayesian inference, combined with generic Monte Carlo algorithms, offers a flexible framework for introducing subjective or context information through the prior, but also, for designing hierarchical models formalizing complex patterns of variation (across sites or branches) or the integration of multiple levels of evolutionary processes. In this chapter, the principles of Bayesian inference, such as applied to phylogenetic reconstruction, are first introduced, with an emphasis on the key features of the Bayesian paradigm that explain its flexibility in terms of model design and its robustness in inferring complex patterns and processes. A more specific focus is then put on the question of modeling pattern-heterogeneity across sites, using both parametric and non-parametric random-effect models. Finally, the current computational challenges are discussed.

How to cite: Nicolas Lartillot (2020). The Bayesian Approach to Molecular Phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 1.4, pp. 1.4:1–1.4:17. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

Bayesian inference was introduced in phylogenetics in the mid 90's (Yang and Rannala, 1997; Mau et al., 1999; Larget and Simon, 1999; Li et al., 2000; Huelsenbeck and Ronquist, 2001). From the very beginnings, it has motivated a lot of discussion about its merits and drawbacks, compared to more firmly established approaches such as maximum likelihood (reviewed by Holder and Lewis (2003); Yang (2007) and in Chapter 1.2 [Stamatakis and Kozlov 2020]). Part of the discussions then revolved around philosophical or foundational issues: how to choose the priors and how to have a control on the sensitivity of the analysis to this choice? What is the meaning of posterior probabilities? How do those compare with alternative measures of statistical support, like non-parametric bootstrap?

Meanwhile, Bayesian inference has reached the stage of practical applications over a broad array of research questions in phylogenomics and in evolutionary studies. To this end, much work has been devoted to the design of increasingly sophisticated models and to the development of efficient algorithms based on Markov Chain Monte Carlo (MCMC). As a result, Bayesian inference, and its relevance to evolutionary genetics, can now be better understood based on its practical impact on current research in our field. As it turns out, the use of Bayesian inference has substantially renewed our perspective on the role of models in phylogenomics (see Chapter 2.1 [Simion et al. 2020]), offering new opportunities that were not directly reachable using classical approaches. Conversely, these practical applications

© Nicolas Lartillot.
Licensed under Creative Commons License CC-BY-NC-ND 4.0.
Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 1.4; pp. 1.4:1–1.4:17
A book completely handled by researchers.
No publisher has been paid.

have shown some weaknesses and limitations of the purely Bayesian philosophical stance, while emphasizing its connections and its similarity with the maximum likelihood paradigm.

The aim of the present chapter is, first, to briefly introduce the basics of Bayesian inference, such as applied to phylogenetics, and to point out its main features in this specific context. In a second step, we will focus on one particular application of Bayesian inference in phylogenomics, namely, the development of Bayesian non-parametric models accounting for variation across sites in amino acid preferences. Finally, we will discuss the current challenges and propose some perspectives concerning how these challenges are being tackled by current statistical and computational research.

2 Bayesian inference in phylogenetics

2.1 General principles of probabilistic inference

Consider a simple phylogenetic problem, in which the aim is to infer the phylogeny of a group of taxa, based on a multiple sequence alignment for a single gene (e.g. ribosomal RNA), and assuming a simple one-parameter process of sequence evolution, the Kimura model (which depends only on the transition-tranversion rate ratio, see Chapter 1.1 [Pupko and Mayrose 2020]). Let us denote by D the sequence alignment, T the unknown tree topology, with branch lengths noted l, and κ the transition/transversion rate ratio. The likelihood is defined as the probability that the sequence evolutionary process with transition-transversion rate ratio κ, running along tree T with branch lengths l, produces the nucleotide sequences D at the tips of the tree. This probability can be written:

$$L(T, l, \kappa) \quad = \quad p(D \mid T, l, \kappa).$$

In the present case, sites are assumed to evolve independently of each other. As a result, the likelihood can be written as a product over all sites. If D_i stands for the site pattern observed at position i, for $i = 1 \ldots n$, where n is the number of aligned positions, then:

$$p(D \mid T, l, \kappa) = \prod_{i=1}^{n} p(D_i \mid T, l, \kappa).$$

The maximum likelihood approach to tree estimation works as follows. First, for a given tree T, the likelihood is maximized with respect to the branch lengths and the parameters of the substitution process:

$$\hat{L}(T) \quad = \quad \max_{l,\kappa} L(T, l, \kappa).$$

This defines a likelihood score for a given tree topology T. Then, we search for the tree topology T which maximizes this score, i.e. the aim is to find the tree $\hat{T}$ such that:

$$\hat{L}(\hat{T}) \quad = \quad \max_{T} \hat{L}(T).$$

Of note, the likelihood is jointly maximized with respect to all unknowns, those we are interested in (here, the tree topology T) and those that are not of direct interest but have an influence on the probability of producing the data (nuisance parameters, here, l and κ). As a result, a tree topology is scored based only on the best-case scenario for all unknown aspects of the evolutionary process under this tree topology, without any consideration for alternative configurations of these unknown nuisances. This is an important point that distinguishes maximum likelihood and Bayesian inference, as we will now see.

2.2 The Bayesian approach

How does Bayesian inference proceed in the present case? First, one has to define a prior distribution over the tree topologies, the branch lengths and the parameter of the substitution model. These priors can be denoted as $p(T)$, $p(l)$ and $p(\kappa)$. Of note, $p(T)$ is a probability over the discrete space of tree topologies. On the other hand, since l and θ are continuous parameters, $p(l)$ and $p(\theta)$ are not probabilities, but probability densities. Which priors have more specifically been used in practical applications will be discussed below. In the present case, one would typically assume simple priors like the following: a uniform prior over T, an exponential of mean 0.1 for branch lengths, and a uniform prior between 0 and 20 for the transition-transversion rate ratio κ.

Second, for a given tree topology T, the likelihood is averaged over all possible values for the continuous parameters l and κ, weighted by the prior distributions over these two parameter components. This defines the marginal likelihood of tree topology T:

$$p(D \mid T) \quad = \quad \int_\theta \int_l p(D \mid T, l, \kappa) p(l) p(\kappa) dl d\kappa. \tag{1}$$

Finally, using Bayes theorem, we can define the posterior probability of the tree topology T:

$$p(T \mid D) \quad = \quad \frac{p(D \mid T)\, p(T)}{p(D)},$$

where the denominator $P(D)$ is the marginal probability of the data (or marginal likelihood), obtained by summing the numerator over all possibles tree topologies (so as to normalize the posterior probability distribution):

$$p(D) \quad = \quad \sum_T p(D \mid T)\, p(T).$$

In many situations, the normalization factor $p(D)$ is not essential, and a simpler account of Bayes theorem is then given by:

$$p(T \mid D) \quad \propto \quad p(D \mid T)\, p(T), \tag{2}$$

which essentially states that the posterior probability of a tree T is proportional to its prior probability $p(T)$ multiplied by the weight of evidence contributed by the data, $p(D \mid T)$. Thus, if $p(T)$ represents our state of belief about which trees are likely to be correct before we have seen the data, then Bayes theorem formalizes how one would update our state of belief upon seeing data D, leading to our posterior state of belief $p(T \mid D)$. For very large datasets, the posterior will typically be concentrated on one single tree topology. In terms of beliefs, this amounts to a nearly complete certainty about that tree being the correct phylogeny, given the data and the model. For smaller datasets, on the other hand, multiple trees may each receive a significant proportion of the total posterior probability mass, which then represents our remaining uncertainty about the phylogenetic history of our clade of interest, given the data.

The equations above could be equivalently rewritten, first by defining a joint posterior over the tree topology and the continuous parameters, using Bayes theorem:

$$p(T, l, \theta \mid D) \quad \propto \quad p(D \mid T, l, \kappa)\, p(T)\, p(l)\, p(\kappa). \tag{3}$$

This formulation is more general, as it puts all unknowns, tree topology, branch lengths, parameters of the sequence evolutionary process, on the same footing. Marginalization is done only in a second step, in a way that depends on the question being asked. Thus far, we

have assumed that the topology was of interest, and thus, we have focussed on the marginal posterior on T, which is obtained by integrating out l and κ:

$$p(T \mid D) \quad = \quad \int_{\kappa} \int_{l} p(T, l, \kappa \mid D) dl d\kappa. \tag{4}$$

Of note, this definition of the marginal posterior on T is equivalent to that given by Eq. 2 above. If instead we were interested in estimating the transition/transversion rate ratio κ, then we would consider the marginal posterior distribution over κ, which is summed over all possible tree topologies and branch lengths:

$$p(\kappa \mid D) \quad = \quad \sum_{T} \int_{l} p(T, l, \kappa \mid D) dl. \tag{5}$$

In the end, Bayesian inference always reduces to computing the posterior probability of what we want to know, given the available evidence and the structural assumptions of the generating model, and this, integrated (averaged) over all other unknowns.

2.3 Practical Bayesian inference using Monte Carlo

The equations indicated above involve sums over a large number of alternative tree topologies and, for a given topology T, integrals over all possible branch lengths and all values for κ. In general, those integrals are not analytically available and would be difficult to numerically evaluate with sufficient accuracy. As a practical alternative, Bayesian inference is most often implemented using Monte Carlo approaches.

The general aim of Monte Carlo is to design random sampling algorithms targeting a probability distribution of interest – here, the joint posterior distribution (Eq. 3). By far the most commonly used algorithm is the Markov chain Monte Carlo (MCMC) approach. The idea of MCMC is to implement a random walk in the parameter space of the model, such that parameter configurations are visited at a frequency proportional to their posterior probability. Running the algorithm for a sufficiently long time yields samples from the posterior distribution:

$$(T_j, l_j, \kappa_j) \quad \sim \quad p(T, l, \kappa \mid D)$$

for $j = 1 \ldots N$, with N a suitably large number of samples. In the case of MCMC, the samples are not independent (successive samples are typically correlated). Furthermore, they are from the targeted posterior distribution only asymptotically. In practice, this means that, starting from an arbitrary parameter configuration, the chain reaches its stationary state only after a burn-in period, and it is only after this stationary state has been reached that samples can be considered to be representative draws from the posterior.

Once such a large sample has been obtained, any marginal over the posterior probability can then be approximated simply by averaging over this sample. For instance, the frequency of a given tree topology T in the sample will be a Monte Carlo estimate of the posterior probability $p(T \mid D)$. More generally, the frequency at which a given group of taxa is found monophyletic in the sample is a Monte Carlo estimate of the posterior probability that this clade is monophyletic. Accordingly, a convenient way to summarize the analysis is to draw the majority-rule consensus of all trees collected by Monte Carlo and label each clade with the Monte Carlo estimate of its posterior probability support.

If, on the other hand, our interest is in some continuous parameters, say the transition/transversion rate ratio, then the histogram of the values of κ collected in the Monte Carlo represents our estimate of the posterior probability density over this parameter. From

there, a 95% credible interval can be computed, by sorting the samples by increasing value and excluding the 2.5% most extreme values at both ends.

2.4 Some important properties of Bayesian inference

Based on the general description of Bayesian inference given above, several points are worth pointing out and discussing.

Averaging over uncertainty

First, Bayesian inference, when conducted on some parameter of interest, always averages uncertainty over all other nuisance parameters. This has already been emphasized above (Eqs. 4 and 5), but some intuition can also be gained from how the output of the MCMC is processed. For instance, as was just pointed out, the Monte Carlo estimate of the posterior probability of a given clade is just the frequency at which this clade is present in the trees sampled by Monte Carlo. Importantly, all trees presenting this specific clade might differ from each other in many other respects – in terms of branch lengths, parameter values, but also in terms of the other clades that are present elsewhere in the tree. By this, we see that, in our evaluation of how likely it is that a given group is monophyletic, we have averaged our evaluation over many possible outcomes for all other aspects of the problem – in some sense, we have diversified our inference portfolio. This is in sharp contrast with maximum likelihood, which bets on one single configuration for the nuisance parameters when deciding for its point estimate. Averaging over uncertainty is expected to lead to more robust inference, in particular, when there are many nuisance parameters (Huelsenbeck et al., 2000).

Monte Carlo versus analytical integration

The point just discussed also shows a key relation between Monte Carlo and integration. Namely, the simple fact of sampling from the joint posterior over tree topology and other parameters and then discarding all parameters, keeping only the tree topology, is equivalent to sampling tree topologies from their marginal posterior distribution. In other words, Monte Carlo automatically implements Eq. 4, and this, without ever explicitly calculating this integral. This apparently anecdotal mathematical observation has important practical consequences: whenever a likelihood is a complex integral over many nuisance variables, then, instead of explicitly calculating this integral, it is always possible to explicitly sample from the nuisance variables, jointly with the parameter of interest. This approach of parameter expansion (or data augmentation) makes it possible to implement a broad category of models that would otherwise not be accessible by explicit numerical integration.

Priors

As mentioned above, averaging the likelihood over the nuisance parameters is expected to lead to more robust inference. On the other hand, this average is taken over a specific prior. More generally making decisions based on posterior probabilities, which are directly proportional to the prior, raises the question of how to choose the priors and how the inference will depend on this choice. Prior choice is perhaps the most important question in Bayesian inference, with many consequences, both conceptual and practical. There is a vast literature on the question, and many questions are still open. The problem is complex, since there are in fact very different approaches to prior definition, proceeding from different philosophies and resulting in posterior probabilities that do not have the same operational properties or the

same meaning. As a tentative typology, it may be useful to distinguish between the following approaches to prior elicitation:

- Informative priors based on expert knowledge. These priors are typically advocated by the so-called subjective Bayesian school, initiated by De Finetti (1974). These priors are not so often used in phylogenetics, with the important exception of soft fossil calibrations in molecular dating (see Yang and Rannala (2006); Chapter 5.1 [Pett and Heath 2020]).
- Uninformative, default or reference priors. The various names given to these priors express their slightly different objectives (priors expressing lack of information, meant to be used by default when expert knowledge is not available, or providing a neutral reference for summarizing empirical information), but in the end, all converge toward very similar practical recommendations. These priors define what is sometimes referred to as the objective Bayesian philosophy (Berger, 2006). In practice, many priors used in Bayesian phylogenetics are meant to be uninformative, or at least sufficiently vaguely informative, so that they can be used by default (without invoking case-dependent expertise or prior information). For instance, the most widely used prior for reconstructing phylogenies in an undated context is a uniform prior over all unrooted tree topologies (Huelsenbeck and Ronquist, 2001) .
- Hierarchical priors. These priors correspond to the closely related empirical or hierarchical Bayesian philosophies. A good example in phylogenetics is to allow for uncorrelated gamma distributed rates across branches in a relaxed clock analysis (Lepage et al., 2007). In this context, branch-specific substitution rates share the same gamma prior. In turn, the shape and scale parameters of this gamma prior, which tune the mean and the variance of rates across branches, are also unknown. Accordingly, a second-stage prior is invoked over these two hyper-parameters. This hyper-prior is typically chosen to be vaguely informative, and as a result, the posterior distribution on the shape and scale parameters will be mostly dictated by the signal about rate variation contained in the sequence data. Hierarchical priors thus represent a powerful tool for designing models allowing for complex modulations of the evolutionary process across branches, sites, or genes, in a way that will be automatically tuned to the true amount of variation present in the data.
- Mechanistic priors, i.e. priors that are themselves justified on the grounds of some macro-evolutionary mechanism or process. A good example is the birth-death prior over phylogenetic trees in a dated context (Yang and Rannala, 1997), which essentially implements a model of species diversification with constant speciation and extinction rates. In a sense, mechanistic priors proceed from the realization that our prior knowledge for some parameter of our problem (here, the unknown phylogeny) is best formalized in terms of a generating model – thus not unlike the likelihood itself. In the end, the result is not very different from the mixed models typically considered in a maximum likelihood context.

Flexibility in model design

The use of mechanistically inspired and hierarchical priors, combined with generic Monte Carlo approaches, has an important practical consequence. Indeed, it makes it possible to design hierarchical models, articulating together multiple levels of processes and integrating multiples sources of empirical data. In this direction, many important developments have taken place over the last decade, including the following:

- Brownian processes for relaxing the molecular clock (see Thorne and Kishino (2002); Chapter 4.4 [Bromham 2020]);

- birth-death models for describing speciation-extinction-fossilization processes (see Heath et al. (2014); Chapter 5.1 [Pett and Heath 2020]);
- integration of the comparative method (models of trait evolution) with the relaxed molecular clock (Lartillot and Poujol, 2011);
- integration of morphological and genetic sequence data (total-evidence dating) (Zhang et al., 2016)
- explicit models of gene duplication, loss and horizontal transfer over species trees (see Akerborg et al. (2009); Chapter 3.2 [Boussau and Scornavacca 2020]);
- multi-species coalescent approaches (see Yang and Rannala (2010); Heled and Drummond (2010); Chapter 3.3 [Rannala et al. 2020]);
- priors over viral phylogenies derived from epidemiological models (see (Kühnert et al., 2014); Chapter 5.3 [Zhukova et al. 2020]);
- non-parametric models for modeling random-effects across sites (Lartillot and Philippe, 2004; Huelsenbeck and Suchard, 2007) or branches (Heath et al., 2012).

The list is not exhaustive, and we can expect many new developments along similar lines. Some of these integrative or hierarchical models are introduced in other chapters of this book (see references above).

3 Bayesian non-parametric site-heterogeneous models

3.1 Variation across sites

One of the most prominent features of multiple sequence alignments at large evolutionary scale is the amount of variation across sites in the degree and the patterns of conservation. This can be seen both at the nucleotide and the amino acid levels. The reasons for this are well understood: sequences that are conserved over long evolutionary periods are almost certainly under strong purifying selection. Selection, however, is highly context-specific, and is thus likely to be fairly disparate across sites of a gene, both in overall intensity and in the nature of the preferred nucleotides or amino acids. This problem is further amplified by the fact that phylogenetic reconstruction is normally conducted using those genes and gene regions that can be reliably aligned over a broad phylogenetic scale. Selecting well-aligned sequences is essential, in order to guarantee the validity of the assumption that sequence variation in the data matrix is only caused by point substitutions, an assumption made by virtually all models currently used in phylogenetics. However, doing so induces a selection bias for those genes and gene regions whose structure is highly conserved. In turn, this means that a given column of the alignment corresponds to a site in the protein (or in the ribosomal RNA) sitting in a very specific biochemical environment inducing strong site-specific purifying selection for maintaining the conformational stability of the macromolecule (e.g. buried sites will accept only hydrophobic amino acids, exposed sites polar amino acids, etc), selection which is stable in the long run (Ashenberg et al., 2013).

In terms of the resulting sequence evolutionary process, this modulation of the selective constraint across sites will translate into a variation in both the rate and the patterns of substitutions across aligned positions. How will such a widespread variation impact phylogenetic estimation? Should we explicitly account for this variation across sites in the model used for phylogenetic reconstruction, or is it sufficient to capture the average substitution process across all sites? If explicit modeling turns out to be important for phylogenetic accuracy, then, how can we design models that will accurately capture the distribution of substitution rates and patterns across sites? These have been important questions in recent phylogenomics.

3.2 Variation of rates across sites: a parametric random-effect model

Accounting for heterogeneity in rates across sites was proposed early on, in a maximum likelihood context (see Yang (1994); Chapter 1.1 [Pupko and Mayrose 2020]). The parametric random-effect approach that was then used was subsequently ported to Bayesian inference, without major modification. It may be useful to look in detail at the conceptual structure of this approach, before addressing the more challenging question of how to model pattern-heterogeneity across sites.

The fundamental idea of the rates-across-sites model introduced by Yang (1994) is to consider site-specific relative rates as random variables, whose distribution across sites is assumed to be a gamma, of mean 1 (since these rates are relative), and of unknown variance tuned by a shape parameter noted α. Mathematically, the likelihood at site i is thus integrated over all possible values for the rate of evolution r, over a gamma distribution:

$$p(D_i \mid \theta, \alpha) \quad = \quad \int_r p(D_i \mid \theta, r) f_\alpha(r) dr,$$

where $f_\alpha(r)$ is the probability density function of the gamma distribution and, for notational simplicity, we refer to all global parameters of the model (other than α) by θ (tree topology, branch lengths, and parameters of the model of sequence evolution). In practice, this integral is intractable, and the standard approach is to numerically estimate it by discretization over a small number K of rate values $(r_k)_{k=1\ldots K}$ centered on the K quantiles of the gamma distribution (typically $K = 4$):

$$p(D_i \mid \theta, \alpha) \quad \simeq \quad \frac{1}{K} \sum_{k=1}^{K} p(D_i \mid \theta, r_k). \tag{6}$$

Finally, we can take the product over all sites:

$$p(D \mid \theta, \alpha) = \prod_i p(D_i \mid \theta, \alpha)$$

which gives the likelihood for the whole sequence alignment. This likelihood can be maximixed with respect to θ and α. Alternatively, in a Bayesian settings, one would define priors over the parameters of the model, θ and α, and then sample from the joint posterior distribution:

$$p(\theta, \alpha \mid D) \quad \propto \quad p(D \mid \theta, \alpha) p(\theta) p(\alpha).$$

Some comments and precisions are in order. First, site-specific rates are integrated over a distribution, and the distribution itself (specifically, its variance, which is equal to $1/\alpha$) is estimated across sites. In a maximum likelihood context, an alternative approach would be possible, at least in principle, namely, maximizing the likelihood with respect to all rates at all sites. However, unless the number of taxa is very large and the tree very long, this would result in overfitting. The site-specific rates would be estimated with large stochastic errors, which would then induce further estimation error on the tree topology. In addition, the variance in the rates thus estimated across sites would be greater than the variance of the true rates, since it would include the additional contribution of the error on rate estimation.

In contrast, dealing with rates as random-effects automatically discounts this additional sampling variance and returns, through α, a fair estimate of the variance of the true rates across sites. This random-effect model also achieves a higher accuracy in the estimation of the tree topology and the global parameters. This is an important point about random-effect models more generally: after fitting the model to the data, we may still have a large uncertainty about the value of the random-effects, yet, in many situations, we will nevetheless

achieve asymptotically consistent estimation of their *distribution*, and of the global parameters of the model (in particular, the tree topology). This is a recurrent idea in many other settings (e.g. integrating over gene genealogies in the multi-species coalescent [Yang 2002]).

Second, the gamma rates model considered above is a *parametric* random-effect model, since we make the assumption that the distribution of rates across sites belongs to a parametric family which is specified in advance (here, a gamma distribution). Nothing guarantees that this assumption will not be violated in practice. The true distribution of rates across sites could be arbitrary, and a mismatch between this true distribution and the distribution assumed by the model could in principle have a non-negligible impact on the accuracy of the estimation of the phylogeny. In general, it is commonly assumed that a gamma (or a mix of a proportion of invariant sites and a gamma distribution) will provide a good enough description of the true distribution of rates across sites, at least for the purpose of phylogenetic reconstruction. Of note, alternative approaches have been proposed to relax this assumption specifically for rates (Mayrose et al., 2005; Huelsenbeck and Suchard, 2007), some of which are similar to those considered below in the case of pattern-heterogeneity.

3.3 Amino acid preferences across sites: non-parametric models

As mentioned above, not just rates, but nucleotide substitution or amino acid replacement patterns, may vary across sites. In the following, and for the sake of the argument, we will more specifically consider the case of amino acid sequences. Given that the primary factor that varies across sites is selection, perhaps the most important feature whose variation across sites should be modeled is amino acid preferences, as a proxy for amino acid fitness. Mathematically, site-specific amino acid preferences can be captured through the 20-dimensional vector of amino acid equilibrium frequencies of the process. In the following, this vector will be called an amino acid frequency *profile*.

An analogy with site-specific rates suggests that we should model amino acid profiles as site-specific random effects, and also, that we should have a method allowing for a sufficiently accurate estimation of the true distribution of amino acid profiles across sites. However, there are important technical differences between site-specific rates and site-specific amino acid profiles, which are such that the method used for rates cannot be directly generalized to the present context. First, the quantile-based discretization approach mentioned above for integrating the likelihood over site-specific rates (Eq. 6) does not scale up well to higher-dimensional random-effects and would not work in practice for 20-dimensional frequency vectors. Another problem is that the true distribution of amino acid profiles is potentially complex, possibly multimodal, and thus probably not well described by any known simple parametric distribution.

Mixture models

A possible alternative is to use a finite mixture model (Koshi and Goldstein, 1998; Pagel and Meade, 2004). The rationale behind mixture models is that the diversity of the patterns of amino acid preferences realized across the aligned positions of empirical sequences might hopefully be captured by a reasonably small number of typical amino acid profiles (e.g. hydrophobic, polar, negatively charged, aromatic, etc). Allowing for K components, each with its own 20-dimensional frequency profile π_k and its own weight w_k, the likelihood at site i is then a weighted average over all mixture components:

$$p(D_i \mid \theta, \pi, w) \simeq \sum_{k=1}^{K} w_k p(D_i \mid \theta, \pi_k) \tag{7}$$

and then taking the product over all sites:

$$p(D \mid \theta, \pi, w) = \prod_{i=1}^{n} p(D_i \mid \theta, \pi, w)$$

gives the likelihood, which now depends on the set of profiles (collectively noted π) and the weight vector w, in addition to the other parameters of the model, collectively referred to as θ. In a maximum likelihood context, this likelihood will typically be maximized with respect to θ, π and w. Alternatively, the series of K profiles can be pre-estimated on a database and then kept fixed during phylogenetic inference (so-called empirical mixture models).

Much effort has been spent on deriving empirical mixtures that could be routinely used in phylogenetics (Quang et al., 2008; Le et al., 2008, 2012; Wang et al., 2014). Thus far, however, this approach has produced mixed results. One main problem is that the number of distinct profiles that seems to be required in order to obtain a good empirical fit and a sufficient phylogenetic accuracy is high (Quang et al., 2008), suggesting that the true distribution of amino acid preferences across sites might be too complex, or too diffuse, to be described by a small number of typical amino acid profiles. Practically, however, allowing for a large number of components quickly raises computational and statistical challenges, at least in a maximum likelihood framework. Computationally, averaging the likelihood at each site over all profiles of the mixture (Eq. 7) becomes prohibitive for large K. Statistically, rich mixtures quickly become redundant, in the sense that many alternative mixture configurations, differing only in small details (e.g. with several components having similar profiles), will typically give essentially equivalent approximations of the unknown empirical distribution, thus leading to poorly identifiable models.

These problems, however, are not so critical in a Bayesian framework, for two different reasons, related to the way Bayesian inference deals with model complexity (see above, Section 2). First, in a Bayesian MCMC context, parameter expansion can be used to avoid the explicit sum over all components for each site (Eq. 7). Instead, one can explicitly sample the allocations of sites to the components of the mixture during the MCMC. Combining this approach with various data-augmentation strategies allows one to design an MCMC strategy whose complexity becomes relatively insensitive to the number of components of the mixture (Lartillot, 2006). Second, the redundancy of rich mixtures, i.e. the fact that alternative mixtures effectively emulate the same distribution of random-effects, is automatically taken care of by averaging out over the posterior distribution of all possible mixture configurations.

Non-parametric random-effect models

These observations suggest that we can in fact use mixture models in a completely different regime: rather than trying to keep the number of components as low as possible, at the cost of not correctly capturing the true empirical diversity of biochemical profiles, one can instead aim for very rich and redundant mixtures. Doing so gives more flexibility. Sufficiently rich mixtures can approximate any distribution with arbitrary accuracy, and the fact that they are redundant does not matter so much, as long as an efficient MCMC is able to smooth out this redundancy by averaging over a representative sample of alternative mixture configurations, all of which giving essentially equivalent approximations of the true distribution. This is the fundamental idea behind Bayesian non-parametric random-effect models.

The original goal of non-parametric inference is to relax the assumption that the true distribution should a priori belong to a pre-specified parametric family. In principle, a non-parametric approach should give asymptotically consistent results for arbitrary distributions of random effects across sites. In a Bayesian context, this is implemented by designing a prior

over rich mixtures. The Dirichlet process is such a non-parametric prior (Ferguson, 1973; Müller and Mitra, 2013). Technically, the Dirichlet process pushes the idea of sufficiently rich mixtures to its extreme, by implementing a prior over *infinite* mixtures. Infinite mixtures are dense in the space of all possible distributions, and thus, a Dirichlet process prior will put some probability mass in the vicinity of any distribution – including of course the true distribution of random-effects across sites. Then, conditioning the model on a sufficiently large dataset will result in a posterior distribution which will concentrate in the vicinity of the true distribution. In the end, implementing this idea using clever MCMC approaches based on parameter expansion will effectively implement a powerful non-parametric inference method, in principle achieving asymptotic consistency under arbitrary distributions of (possibly multi-dimensional) random-effects.

Dirichlet process priors have been applied to several problems in phylogenetics, for modeling variation across sites in rates (Huelsenbeck and Suchard, 2007), dN/dS (Huelsenbeck et al., 2006), amino acid preferences across sites (Lartillot and Philippe, 2004) or amino acid fitness profiles in the context of mechanistic mutation-selection codon models (Rodrigue et al., 2010); but also, for modeling variation in rates across branches in a relaxed clock model (Heath et al., 2012).

4 Software programs and platforms

A large number of software programs are currently available for conducting phylogenetic or phylogeny-related Bayesian inference. These programs often have very different specific objectives or specializations: phylogenetic reconstruction (Ronquist and Huelsenbeck (2003); Lartillot et al. (2013); Lewis et al. (2015); Chapter 1.5 [Lartillot 2020]), molecular dating (Thorne and Kishino (2002); Rannala and Yang (2007); Chapter 5.2 [Barido-Sottani et al. 2020]), phylogeography and phylodynamics (Bouckaert et al., 2014), phylogenetic codon models (Murrell et al. (2013); Rodrigue and Lartillot (2014); Chapter 4.5 [Lowe and Rodrigue 2020]), comparative studies (Pagel et al., 2004), gene-tree species-tree reconciliation (Akerborg et al., 2009), or species delimitation (Yang and Rannala (2010); Chapter 5.6 [Flouri et al. 2020]).

In spite of this current move toward integrative modeling approaches (and perhaps in part because of the computational challenges), much of current applied research in phylogenomics is still conducted in the context of the more classical supermatrix paradigm, in which a large set of single-gene alignments are simultaneously considered and assumed to evolve along the same species tree. Three main software programs have been used in recent phylogenomic analyses more specifically devoted to reconstructing a species tree using supermatrices:

- MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) is the most widely used program for Bayesian phylogenetic reconstruction. It offers a broad range of models, allowing for standard nucleotide, amino acid and codon models, but also models of morphological character evolution, while offering the possibility to analyze heterogeneous datasets (i.e. mixing morphological, DNA or RNA and amino acid data). On the other hand, this program has limited expressivity for pattern-heterogeneity across sites within partitions. MrBayes has been extensively used for standard phylogenetic analysis, and in several recent phylogenomic studies (Cannon et al., 2016).
- PhyloBayes (Lartillot et al., 2009, 2013) is a program specialized in site-heterogeneous models of sequence evolution. Its main distinguishing feature is the use of Dirichlet process priors, such as introduced above, to model the variation across sites in amino acid preferences. In part because of the increasing awareness of the importance of accounting

for site-heterogeneity for reconstructing ancient phylogenies, this program has been increasingly used over the recent years, in particular for reconstructing the metazoan tree of life (Simion et al., 2017), but also eukaryotes (Brown et al., 2018), Archaea (Adam et al., 2017), or Eubacteria (Antunes et al., 2016). A detailed application using PhyloBayes is presented in Chapter 1.5 of this book (Lartillot, 2020).

- ExaBayes (Aberer et al., 2014) is a recent implementation of Bayesian phylogenetic inference for very large supermatrices. This program allows for heterogeneity across partitions, as well as rate-heterogeneity (but not pattern-heterogeneity) across sites within partitions.
- P4 (Foster, 2004) is specialized in branch-heterogeneous models of sequence evolution, more specifically, accounting for compositional heterogeneity across taxa. This program has been used, in particular, for investigating the position of eukaryotes in the tree of life (Cox et al., 2008).

Ideally, it would be very useful to have a single implementation combining these various levels of expressiveness in model design (i.e. allowing for gene-, site- and branch-specific modulations in both rate and pattern heterogeneity), all of which appear to be essential in order to achieve accurate phylogenetic reconstruction. Thus far, however, no such integrated implementation is available. One main reason for this vacancy is the computational complexity inherent to each of these multiple sources of variation, which would be compounded in a joint implementation and further aggravated by the size of the current datasets in phylogenomics.

5 Conclusions and perspectives

In several respects, Bayesian inference has revolutionized our practice in phylogenetics, although perhaps not for the reasons that have often been invoked. In theory, Bayesian inference offers a flexible framework for introducing subjective or context information through the prior. In practice, however, this is not the main reason behind the recent success and popularity of Bayesian inference in evolutionary genetics. Instead, it is the combination of hierarchical models and generic Monte Carlo approaches for dealing with complex random-effects and multi-level evolutionary processes that has played the most important role.

Modeling pattern-heterogeneity across sites represents one specific instance where complex random-effect models turn out to have an important impact on practical phylogenomics. This problem is challenging in two respects: first, because the random effects to be modeled (amino acid preferences) are high-dimensional, and second, because the distribution of those random-effects across sites is itself unknown and apparently complex. This combination makes Bayesian inference using Monte Carlo particularly well-suited, whereas simpler approaches, such as parametric random-effect models or finite mixture models, have thus far shown less affordable or less accurate.

However, at least given the current state-of-the art, Bayesian inference suffer from several limitations. First, current Monte Carlo algorithms do not scale up well with data size, and as a result, Bayesian inference can become computationally prohibitive for large phylogenomic datasets. This is particularly true for non-parametric models based on Dirichlet process priors. Second, the flexibility afforded by Bayesian inference for handcrafting new and complex multi-level models is nice in theory, yet in practice, it requires a substantial amount of programming work for each new model that one might want to contemplate. This also raises the question of the reliability of the software implementations, as it is typically difficult to guarantee that a given implementation of a Monte Carlo algorithm is indeed sampling from the intended target distribution.

In this direction, the current trend is in the development of generic programming platforms. The fact that model components can be composed like building blocks into complex hierarchical structures, using modular Monte Carlo methods to explore the resulting posterior distribution, makes generic programming for Bayesian inference relatively straightforward, at least conceptually. Such generic programming platforms have been proposed both in general applied statistics (Lunn et al., 2009) and more recently in phylogenetics (Höhna et al. 2016; Chapter 5.4 [Ayres et al. 2020]). They represent a promising development, by providing the applied evolutionary scientific community with user-oriented tools for reliably designing question-specific integrative models and applying them to arbitrary combinations of empirical data (genetic sequences, morphological data, time series, etc). The computational challenges, however, are formidable, and much remains to be done in Monte Carlo algorithmics and software development, in order for Bayesian generic programming to achieve scalable inference in the context of current problems in evolutionary genomics.

References

Aberer, A. J., Kobert, K., and Stamatakis, A. (2014). ExaBayes: massively parallel bayesian tree inference for the whole-genome era. *Mol. Biol. Evol.*, 31(10):2553–2556.

Adam, P. S., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. (2017). The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J*, 11(11):2407–2425.

Akerborg, O., Sennblad, B., Arvestad, L., and Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. USA*, 106(14):5714–5719.

Antunes, L. C., Poppleton, D., Klingl, A., Criscuolo, A., Dupuy, B., Brochier-Armanet, C., Beloin, C., and Gribaldo, S. (2016). Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *Elife*, 5.

Ashenberg, O., Gong, L. I., and Bloom, J. D. (2013). Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl. Acad. Sci. USA*, 110(52):21071–21076.

Ayres, D. L., Lemey, P., Baele, G., and Suchard, M. A. (2020). Beagle 3 high-performance computational library for phylogenetic inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.4, pages 5.4:1–5.4:9. No commercial publisher | Authors open access book.

Barido-Sottani, J., Justison, J. A., Wright, A. M., Warnock, R. C. M., and Pett, W. (2020). Estimating a time-calibrated phylogeny of fossil and extant taxa using revbayes. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.2, pages 5.2:1–5.2:23. No commercial publisher | Authors open access book.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.*, 10(4):e1003537.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Bromham, L. (2020). Substitution rate analysis and molecular evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.4, pages 4.4:1–4.4:21. No commercial publisher | Authors open access book.

Brown, M. W., Heiss, A. A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A. K., Shiratori, T., Ishida, K.-I., Hashimoto, T., Simpson, A. G. B., and Roger, A. J. (2018). Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biol. Evol.*, 10(2):427–433.

Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, 530(7588):89–93.

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., and Embley, T. M. (2008). The archaebacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. USA*, 105(51):20356–20361.

De Finetti, B. (1974). *Theory of Probability.* John Wiley and Sons, New York.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230.

Flouri, T., Rannala, B., and Yang, Z. (2020). A tutorial on the use of bpp for species tree estimation and species delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.6, pages 5.6:1–5.6:16. No commercial publisher | Authors open access book.

Foster, P. G. (2004). Modeling Compositional Heterogeneity. *Syst. Biol.*, 53(3):485–495.

Heath, T. A., Holder, M. T., and Huelsenbeck, J. P. (2012). A dirichlet process prior for estimating lineage-specific substitution rates. *Mol. Biol. Evol.*, 29(3):939–955.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc. Natl. Acad. Sci. USA*, 111(29):E2957–66.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27(3):570–580.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Syst. Biol.*, 65(4):726–736.

Holder, M. and Lewis, P. (2003). Phylogenetic estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.*, 4:275–284.

Huelsenbeck, J., Rannala, B., and Masly, J. (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288(5475):2349–2350.

Huelsenbeck, J. P., Jain, S., Frost, S. W. D., and Pond, S. L. K. (2006). A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl. Acad. Sci. USA*, 103(16):6263–6268.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Huelsenbeck, J. P. and Suchard, M. A. (2007). A nonparametric method for accommodating and testing across-site rate variation. *Syst. Biol.*, 56(6):975–987.

Koshi, J. M. and Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins*, 32(3):289–295.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. (2014). Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *Journal of The Royal Society Interface*, 11(94):20131106.

Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, 16:750–759.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *J. Comput. Biol.*, 13(10):1701–1722.

Lartillot, N. (2020). Phylobayes: Bayesian phylogenetics using site-heterogeneous models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.5, pages 1.5:1–1.5:16. No commercial publisher | Authors open access book.

Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288.

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

Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.*, 28(1):729–744.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, 62(4):611–615.

Le, S. Q., Dang, C. C., and Gascuel, O. (2012). Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.*, 29(10):2921–2936.

Le, S. Q., Lartillot, N., and Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 363(1512):3965–3976.

Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. (2007). A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.*, 24(12):2669–2680.

Lewis, P. O., Holder, M. T., and Swofford, D. L. (2015). Phycas: software for Bayesian phylogenetic analysis. *Syst. Biol.*, 64(3):525–531.

Li, S., Pearl, D. K., and Doss, H. (2000). Phylogenetic Tree Construction Using Markov Chain Monte Carlo. *J. Am. Stat. Assoc.*, 95(450):493–508.

Lowe, C. and Rodrigue, N. (2020). Detecting adaptation from multi-species protein-coding dna sequence alignments alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.5, pages 4.5:1–4.5:18. No commercial publisher | Authors open access book.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

Mau, B., Newton, M. A., and Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55(1):1–12.

Mayrose, I., Friedman, N., and Pupko, T. (2005). A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, 21 Suppl 2:ii151–8.

Müller, P. and Mitra, R. (2013). Bayesian Nonparametric Inference - Why and How. *Bayesian Analysis*, 8(2).

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., and Scheffler, K. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.*, 30(5):1196–1205.

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, 53(4):571–581.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.*, 53(5):673–684.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Quang, L. S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Rannala, B. and Yang, Z. (2007). Inferring speciation times under an episodic molecular clock. *Syst. Biol.*, 56(3):453–466.

Rodrigue, N. and Lartillot, N. (2014). Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics*, 30(7):1020–1021.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. USA*, 107(10):4629–4634.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, E., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., and Manuel, M. (2017). A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol.*, 27(7):958–967.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Thorne, J. L. and Kishino, H. (2002). Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.*, 51(5):689–702.

Wang, H.-C., Susko, E., and Roger, A. J. (2014). An amino acid substitution-selection model adjusts residue fitness to improve phylogenetic estimation. *Mol. Biol. Evol.*, 31(4):779–792.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39(3):306–314.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4):1811–1823.

Yang, Z. (2007). Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. *Mol. Biol. Evol.*, 24(8):1639–1655.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol. Biol. Evol.*, 14(7):717–724.

Yang, Z. and Rannala, B. (2006). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.*, 23(1):212–226.

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. USA*, 107(20):9264–9269.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. (2016). Total-Evidence Dating under the Fossilized Birth-Death Process. *Syst. Biol.*, 65(2):228–249.

Zhukova, A., Gascuel, O., Duchêne, S., Ayres, D. L., Lemey, P., and Baele, G. (2020). Efficiently analysing large viral data sets in computational phylogenomics. In Scornavacca,

C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.3, pages 5.3:1–5.3:43. No commercial publisher | Authors open access book.