Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 683509, 16 pages
doi:10.1155/2008/683509

Review Article

Phylogenetic Analyses: A Toolbox Expanding towards Bayesian Methods

Stéphane Aris-Brosou[1,2] and Xuhua Xia[1]

[1] *Department of Biology, Centre for Advanced Research in Environmental Genomics, University of Ottawa, Ontario, Canada K1N 6N5*
[2] *Department of Mathematics and Statistics, University of Ottawa, Ontario, Canada K1N 6N5*

Correspondence should be addressed to Stéphane Aris-Brosou, sarisbro@uottawa.ca

Received 30 November 2007; Accepted 12 February 2008

Recommended by Chunguang Du

The reconstruction of phylogenies is becoming an increasingly simple activity. This is mainly due to two reasons: the democratization of computing power and the increased availability of sophisticated yet user-friendly software. This review describes some of the latest additions to the phylogenetic toolbox, along with some of their theoretical and practical limitations. It is shown that Bayesian methods are under heavy development, as they offer the possibility to solve a number of long-standing issues and to integrate several steps of the phylogenetic analyses into a single framework. Specific topics include not only phylogenetic reconstruction, but also the comparison of phylogenies, the detection of adaptive evolution, and the estimation of divergence times between species.

1. INTRODUCTION

Human cultures have always been fascinated by their origins as a means to define their position in the world, and to justify their hegemony over the rest of the living world. However, scientific (testable) predictions about our origins had to wait for Darwin [1] and his intellectual descendents first to classify [2] and then to reconstruct the natural history of replicating entities, and hereby to kick-start the field of phylogenetics [3, 4]. Rooted in the comparison of morphological characters, phylogenies have for the past four decades focused on the relationships between molecular sequences (e.g., [4]), potentially helped by incorporating morphological information [5], in order to infer ancestor-to-descendent relationships between sequences, populations, or species.

Today, molecular phylogenies are routinely used to infer gene or genome duplication events [6], recombination [7], horizontal gene transfer [8], variation of selective pressures and adaptive evolution [9], divergence times between species [10], the origin of genetic code [11], elucidate the origin of epidemics [12], and host-parasite cospeciation events [13, 14]. As complementary tools for taxonomy (DNA barcoding: [15]), they have also contributed to the formulation of strategies in conservation biology [16]. In addition to untangling the ancestral relationships relating a group of taxa or of a set of molecular sequences, phylogenies have also been used for some time outside of the realm of biological sciences as for instance in linguistics [17, 18] or in forensics [19, 20].

Most of these applications are beyond the scope of plant genomics, but they all suggest that sophisticated phylogenetic methods are required to address most of today's biological questions. While parsimony-based methods are both intuitive and extremely informative, for instance to disentangle genome rearrangements [21], they also have their limitations due to their minimizing the amount of change [22]. These limitations become particularly apparent when analyzing distantly related taxa. A means to overcome, at least partly, some of these difficulties is to adopt a model-based approach, be in a maximum likelihood or in a Bayesian framework. These two frameworks are extremely similar in that they both rely on probabilistic models. Bayesian approaches offer a variety of benefits when compared to traditional maximum likelihood, such as computing speed (although this is not necessarily true, especially under complex models), sophistication of the model, and an appropriate treatment of uncertainty, in particular the one about nuisance parameters.

As a result, Bayesian approaches often make it possible to address more sophisticated biological questions [23], which usually comes at the expense of longer computing times and higher memory requirements than when using simpler models.

Because it is not possible or even appropriate to discuss all the latest developments in a given field of study, this review will focus on a very limited number of key phylogenetic topics. Of notable exceptions, recent developments in phylogenetic hidden Markov models [24] or applications that map ancestral states on phylogenies [25] are not treated. We focus instead on the very first steps involved in *most* phylogenetic analysis, ranging from reconstructing a tree to estimating selective pressures or species divergence times. For each of these steps, some of the most recent theoretical developments are discussed, and pointers to relevant software are provided.

2. RECONSTRUCTING PHYLOGENETIC TREES

2.1. Sequence alignment

The first step in reconstructing a phylogenetic tree from molecular data is to obtain a multiple sequence alignment (MSA) where sequence data are arranged in a matrix that specifies which residues are homologous [26]. A large number of methods and programs exist [27] and most have been evaluated against alignment databases [28], so that it is possible to provide some general guidelines.

The easiest sequences to align are probably those of protein-coding genes: proteins diverge more slowly than DNA sequences and, as a result, proteins are easier to align. The rule-of-thumb is therefore first to translate DNA to amino acid sequences, then perform the alignment at the protein level, before back-translating to the DNA alignment in a final step. This procedure avoids inserting gaps in the final DNA alignment that are not multiple of three and that would disrupt the reading frame. Translation to amino acid sequences can be done directly when downloading sequences, for instance from the National Center for Biotechnology Information (NCBI: www.ncbi.nlm.nih.gov). A number of programs also allow users to perform this translation locally on their computers from an appropriate translation table (e.g., DAMBE [29], MEGA [30, 31]; see Table 1). The second step is to perform the alignment at the protein level. Again, a number of programs exist, but ProbCons [32] appears to be the most accurate *single* method [33]. An alternative for using one single alignment method is to use consensus or meta-methods, that is, to combine several methods [27]. Meta-methods such as M-Coffee can return better MSAs almost twice as often as ProbCons [34]. Finally, when the alignment is obtained at the protein level, back-translation to the DNA sequences can be performed either by using a program such as DAMBE, CodonAlign [35], or by using a dedicated server such as protal2dna (http://bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html) or Pal2Nal (coot.embl.de/pal2nal).

The alignment of rRNA genes with the constraint of secondary structure has now been frequently used in practical research in molecular evolution and phylogenetics [56–60]. The procedure is first to obtain reliable secondary structure and then use the secondary structure to guide the sequence alignment. This has not been automated so far, although both Clustal [61, 62] and DAMBE have some functions to alleviate the difficulties.

What to do with other noncoding genes is still an open question, especially when it comes to aligning a large number (>100) of long (>20,000 residues) and divergent sequences (<25% identity). Some authors have attempted to provide rough guidelines to choose the most accurate program depending on these parameters [28]. However, accuracy figures are typically estimated over a large number of test alignments and may not reflect the accuracy that is expected for any particular alignment [28]. More crucially, most of the alignment programs were developed and benchmarked on protein data, so that their accuracy is generally unknown for noncoding sequences [28]. A very general recommendation is then to use different methods [63] and meta-methods. Those parts of the alignment that are similar across the different methods are probably reliable. The parts that differ extensively are often simply eliminated from the alignment when no external information can be used to decide which positions are homologous. Poorly aligned regions can cause serious problems as, for instance, when analyzing rRNA sequences in which conserved domain and variable domains have different nucleotide frequencies [60]. A simple test of the reliability of an alignment consists in reversing the orientation of the original sequences, and performing the alignment again; because of the symmetry of the problem, reliable MSAs are expected to be identical whichever direction is used to align the sequences [64]. These authors further show that reliability of MSAs decreases with sequence divergence, and that the chance of reconstructing different phylogenies increases with sequence divergence. More sophisticated methods also permit the direct measure of the accuracy of an alignments or the estimation of a distance between two alignments [65]. Applications of Bayesian inference strictly to pairwise [66] and multiple [67, 68] sequence alignment are still in their infancy.

Whichever method is used to obtain an MSA, a final visual inspection is required, and manual editing is often needed. To this end, a number of editors can be used such as JalView [69].

Because an MSA represents a hypothesis about sitewise homology at all the positions, obtaining an accurate MSA presents some circularity; an accurate MSA often necessitates an accurate guide tree, which in turn demands an accurate alignment. The early realization of this "chicken-egg" conundrum led to the idea that both the MSA and the phylogeny should be estimated simultaneously [70]. Although this initial algorithm was parsimony-based, it was already too complex to analyze more than a half-dozen sequences of 100 sites or more. Subsequent parsimony-based algorithms allowed the analysis of larger data sets [71] but still showed some limitations when sequence divergence increases. More recently, a Bayesian procedure was described and implemented in a program, BAli-Phy, where uncertainties with respect to the alignment, the tree, and the parameters of the substitution

Table 1: *List of programs cited in this review*. GUI: graphic user interface; ML: maximum likelihood; PL: penalized likelihood.

Name	Method	Platform	GUI	Inference	Reference
BAMBE	Bayes	DOS, MacOS, Unix	No	Tree	[36]
BayesPhylogenies	Bayes	DOS, MacOS, Unix	No	Tree	[37]
BAli-Phy	Bayes	DOS, MacOS, Unix	No	Simultaneous alignment and tree	[38]
BEAST	Bayes	Windows, MacOS, Unix	Yes	Tree, times	[39]
CONSEL	ML	DOS, MacOS, Unix	No	Tree comparison	[40]
DAMBE	Distances, parsimony, ML	Windows	Yes	Tree	[29]
GARLI	ML (Genetic Algorithm)	Windows, MacOS, Unix	Yes	Tree	[41]
HyPhy	ML	Windows, MacOS, Unix	Yes	Tree, selection, recombination, tree comparison,	[42]
MEGA	Distances, parsimony	Windows	Yes	Tree, times	[30, 31]
MrBayes	Bayes	DOS, MacOS, Unix	No	Tree, selection	[43, 44]
Multidivtime	Bayes	DOS, MacOS, Unix	No	Times	[45–47]
OmegaMap	Bayes	DOS, MacOS, Unix	No	Simultaneous selection and recombination	[48]
PAML	ML	DOS, MacOS, Unix	No	Tree, tree comparison, times, selection	[49, 50]
PAUP$*$	Distances, parsimony, ML	DOS, MacOS, Unix	No	Tree	[51]
PhyloBayes	Bayes	DOS, MacOS, Unix	No	Tree, tree comparison	[52]
PHYML	ML	DOS, MacOS, Unix	No	Tree	[53]
RAxML	ML	DOS, MacOS, Unix	No	Tree	[54]
r8s	PL	DOS, MacOS, Unix	No	Times	[55]

model are all taken into account [38] (see also [72]). Uncertain alignments are a potential problem in large-scale genomic studies [73] or in whole-genome alignments [74]. In these contexts, disregarding alignment uncertainty can lead to systematic biases when estimating gene trees or inferring adaptive evolution [73, 74]. However, these complex Bayesian models [38, 72, 73] still require some nonnegligible computing time and resource, and to date, their performance in terms of accuracy is still unclear.

2.2. Selection of the substitution model

Once a reliable MSA is obtained, the next step in comparing molecular sequences is to choose a metric to quantify divergence. The most intuitive measure of divergence is simply to count the proportion of differences between two aligned sequences (e.g., [75]). This simple measure is known as the p distance. However, because the size of the state space is finite (four letters for DNA, 20 for amino acids, and 61 for sense codons), multiple changes at a position in the alignment will not be observable, and the p distance will underestimate evolutionary distances even for moderately divergent sequences. This phenomenon is generally referred to as saturation. Corrections were devised early to help compensate for saturation. Some of the most famous named nucleotide substitution models are the Jukes-Cantor model or JC [76], the Kimura two-parameter model or K80 [77], the Hasegawa-Kishino-Yano model or HKY85 [78], the Tamura-Nei model or TN93 [79], and the general time-reversible model or GTR [80] (also called REV). Because substitution rates vary along sequences, two components can be added to these substitution models: a "+I" component that models invariable sites [78] and a "+Γ" component that models among-site rate variation either as a continuous [81] or as a discrete [82] mean-one Γdistribution, the latter being more computationally efficient. Amino acid models can also incorporate a "+F" component so that replacement rates are proportional to the frequencies of both the replaced and resulting residues [83].

Given the variety of substitution models, the first step of any model-based phylogenetic analysis is to select the most appropriate model [84, 85]. The rational for doing so is to balance bias and variance: a highly-parameterized model will describe or fit the data much better than a model that contains a smaller number of parameters; in turn however, each parameter of the highly-parameterized model will be estimated with lower accuracy for a given amount of data (e.g., [86]). Besides, both empirical and simulation studies show that the choice of a wrong substitution model can lead not only to less accurate phylogenetic estimation, but also to inconsistent results [87]. The objective of model selection is therefore not to select the "best-fitting" model, as this one will always be the model with the largest number of parameters, but rather to select the most appropriate model that will achieve the optimal tradeoff between

bias and variance. The approach followed by all model selection procedures is therefore to penalize the likelihood of the parameter-rich model for the additional parameters. Because most of the nucleotide substitution models are nested (all can be seen as a special case of GTR +Γ+I), the standard approach to model selection is to perform hierarchical likelihood ratio tests or hLRTs [88]. Note that in all rigor, likelihood ratio tests can also be performed on nonnested models; however, the asymptotic distribution of the test statistic (twice the difference in log-likelihoods) under the null hypothesis (the two models perform equally well) is complicated [89] and quite often impractical. When models are nested, the asymptotic distribution of the test statistic under the null hypothesis is simply a χ^2 distribution whose degree of freedom is the number of additional parameters entering the more complex model (see [90] or [91] for applicability conditions). With the hLRT, then all models are compared in a pairwise manner, by traversing a choice-tree of possible nested models. A number of popular programs allow users to compare pairs of models manually (e.g., PAUP [51], PAML [49, 50]). Readily written scripts that select the most appropriate model among a list of named models also exist, such as ModelTest [92] (which requires PAUP), the *R* package APE [93], or DAMBE. Free web servers are also available; they are either directly based on ModelTest [94] or implement similar ideas (e.g., FindModel, available at hcv.lanl.gov/content/hcv-db/findmodel/findmodel.html). A similar implementation, ProtTest, exists for protein data [95].

However, performing systematic hLRTs is not the optimal strategy for model selection in phylogenetics [96]. This is because the model that is finally selected can depend on the order in which the pairwise comparisons are performed [97]. The Akaike information criterion (AIC) or its variant developed in the context of regression and time-series analysis in small data sets (AIC_c, [98]) is commonly used in phylogenetics (e.g., [96]). One advantage of AIC is that it allows nonnested models to be compared, and it is easily implemented. However, in large data sets, both the hLRT and the AIC tend to favor parameter-rich models [99]. A slightly different approach was proposed to overcome this selection bias, the Bayesian information criterion (BIC: [99]), which penalizes more strongly parameter-rich models. All these model selection approaches (AIC, AIC_c, and BIC) are available in ModelTest and ProtTest. Other procedures exist such as the Decision-Theoretic or DT approach [100]. Although AIC, BIC, and DT are generally based on sound principles, they can in practice select different substitution models [101]. The reason for doing so is not entirely clear, but it is likely due to the data having low-information content. One prediction is that, when these model selection procedures end up with different conclusions, all the selected models will return phylogenies that are not significantly different. It is also possible that applying these different criteria outside of the theoretical context in which they were developed might lead to unexpected behaviors [102]. For instance, $AIC_c$was derived under Gaussian assumptions for linear fixed-effect models [98], and other bias correction terms exist under different assumptions [86].

All the above test procedures compare ratios of likelihood values penalized for an increase in the dimension of one of the models, without directly accounting for uncertainty in the estimates of model parameters. This may be problematic, in particular for small data sets. The Bayesian approach to model selection, called the Bayes factor, directly incorporates this uncertainty. It is also more intuitive as it directly assesses if the data are more probable under a given model than under a different one (e.g., [103]). An extension of this approach makes it possible to select the model not only among the set of named models (JC to GTR) but among all 203 nucleotide substitution models that are possible [104]. An alternative use or interpretation of this approach is to integrate directly over the uncertainty about the substitution model, so that the estimated phylogeny fully accounts for several kinds of uncertainty: about the substitution models, and the parameters entering each of these models. MrBayes (version 3.1.2) [43] implements this feature for amino acid models.

There is an element of circularity in model selection, just as in sequence alignment. In theory, when the hLRT is used for model selection, the topology used for all the computations should be that of the maximum likelihood tree. In practice, model selection is based on an initial topology obtained by a fast algorithm such as neighbor-joining [105, 106] (default setting in ModelTest) or by Weighbor [107] (default setting in FindModel) on JC distances without any correction for among-site rate variation. As mentioned above, it is known that the choice of a wrong model can affect the tree that is estimated, but it is not always clear how the choice of a nonoptimal topology to select the substitution model affects the tree that is finally estimated. Again, this issue with model choice disappears with Bayesian approaches that integrate over all possible time-reversible models as in [104].

2.3. *Finding the "best" tree and assessing its support*

Once the substitution model is selected, the classical approach proceeds to reconstruct the phylogeny [108]. This is probably one area where phylogenetics has seen mixed progress over the last five years, due to both the combinatorial and the computational complexities of phylogenetic reconstruction.

The combinatorial complexity relates to the extremely large number of tree topologies that are possible with a large number of sequences [109]. For instance, with five sequences, there are 105 rooted topologies, but with ten sequences, this number soars to over 34 million. An exhaustive search for the phylogeny that has the highest probability is therefore not practical even with a moderate number of sequences. Besides, while heuristics exist (e.g., stepwise addition [109]; see [4] for a review), almost none of these is guaranteed to converge on the optimum phylogenetic tree. The common practice is then to use one of these heuristics to find a good starting tree, and then modify repeatedly its topology more or less dramatically to explore its neighborhood for better trees until a stopping rule is satisfied [110]. The art here is in designing efficient tree perturbation methods that adaptively strike a balance between large topological modifications (that almost always lead to a very different tree with

a poor score) and small modifications (that almost always lead to an extremely similar tree with lower score). Some of today's challenges are about choosing between methods that successfully explore large numbers of trees but that can be costly in terms of computing time [110], and methods that are faster but may miss some interesting trees [53]. Several programs such as Leaphy, PhyML, and GARLI[41] are among the best-performing software in a maximum likelihood setting. In a Bayesian framework, the basic perturbation schemes were described early [36] and recently updated [111]. Three popular programs are MrBayes, BAMBE [36], and BEAST [39]. Among all these programs and approaches, PHYML, GARLI, and BEAST are probably among the most efficient programs in terms of computational speed, handling of large data sets and thoroughness of the tree search.

A first aspect of the computational complexity relates to estimating the support of a reconstructed phylogeny. This is more complicated than estimating a confidence interval for a real-valued parameter such as a branch length, because a tree topology is a graph and not a number. The classical approach therefore relies on a nonstandard use of the bootstrap [112]. However, the interpretation of the bootstrap is contentious. Bootstrap proportions P can be perceived as testing the correctness of internal nodes, and failing to do so [113], or $1–P$ can be interpreted as a conservative probability of falsely supporting monophyly [114]. Since bootstrap proportions are either too liberal or too conservative depending on the exact interpretation given to these values [115], it is difficult to adjust the threshold below which monophyly can be confidently ruled out [116]. Alternatively, an intuitive geometric argument was proposed to explain the conservativeness of bootstrap probabilities [117], but the workaround was never actually used in the community or implemented in any popular software. The introduction of Bayesian approaches in the late 1990s [36, 118] suggested a novel approach to estimate phylogenetic support with posterior probabilities. Clade or bipartition posterior probabilities can be relatively fast to compute, even for large data sets analyzed under complicated substitution models [119]. As in model selection, they have a clear interpretation as they measure the probability that a clade is correct, given the data and the model. But as with bootstrap probabilities, some controversies exist. Early empirical studies found that posterior probabilities of highly supported nodes were much larger than bootstrap probabilities [120], and subsequent simulation studies supported this observation (e.g., [121–124]). Some of these differences can be attributed to an artifact of the simulation scheme that was employed [125], but more specific empirical and simulation studies show that prior specifications can dramatically impact posterior probabilities for trees and clades [115, 126, 127]. In the simplest case, the analysis of simulated star trees with four sequences fails to give the expected three unrooted topologies with equal probability (1/3, 1/3, 1/3) but returns large posterior probabilities for an arbitrary topology [115, 126], even when infinitely long sequences are used [128, 129] ([130]). This phenomenon, called the star-tree paradox [126], seems to disappear when polytomies are assigned nonzero prior probabilities and when nonuniform priors force internal branch length towards zero [129]. The second issue surrounding Bayesian phylogenetic methods is about their convergence rate. A theoretical study shows that extremely simple Markov chain Monte Carlo (MCMC) samplers, the technique used to estimate posterior probabilities, could take an extremely long time to converge [131]. In practice, however, MCMC samplers such as those implemented in MrBayes are much more sophisticated. In particular, they include different types of moves [111] and use tempering, where some of the chains of a single run are heated, to improve mixing [43]. As a result, it is unclear whether they suffer from extremely long convergence times. It is also expected that current convergence diagnostic tools such as those implemented in MrBayes would reveal convergence problems [132]. Finally, it is also argued that these controversies such as exaggerated clade support, inconsistently biased priors, and the impossibility of hypothesis testing disappear altogether when posterior probabilities at internal nodes are abandoned in favor of posterior probabilities for topologies [133] (see Section 2.4 below).

The most fundamental aspect of the computational complexity in phylogenetics is due to the structure of the phylogenies: these are trees or binary graphs on which computations are nested and interdependent, which makes these computations intractable or NP-hard [134]. As a result, it is difficult to adopt an efficient "divide and conquer" approach, where a large complicated problem would be split into small simpler tasks, and to take advantage of today's commodity computing by distributing the computation over multicore architectures or heterogeneous computer clusters. Current strategies are limited to distributing the computation of the discrete rate categories (when using a "+Γ" substitution model) and part of the search algorithm [54], or simply to distribute different maximum likelihood bootstrap replicates [53, 54] or different MCMC samplers to available processors [44].

2.4. Comparisons of tree topologies

Science proceeds by testing hypotheses, and it is often necessary to compare phylogenies, for instance to test whether a given data set supports the early divergence of gymnosperms with respect to Gnetales and angiosperms (the anthophyte hypothesis), or whether the Gnetales diverged first (the Gnetales hypothesis) [135, 136]. Because of the importance of comparing phylogenies, a number of tests of molecular phylogenies were developed early. The KH test was first developed to compare two random trees [137]. However, this test is invalid if one of the trees is the maximum likelihood tree [138]. In this case, the SH test should be used [139]. Because the SH test can be very conservative, an approximately unbiased version was developed: the AU test [140]. PAUP and PAML only implement the KH and SH tests; CONSEL [40] also implements the AU test. A Bayesian version of these tests also exists [141], but the computations are more demanding.

Indeed, the Bayesian approach to hypothesis testing relies on computing the probability of the data under a particular model. This quantity is usually not available as a close-form equation, and it must be approximated numerically. The most straightforward approximation is based on the harmonic mean of the likelihood sampled from the posterior

distribution [142]. This approximation was described several times in the context of phylogenies [141, 143] and is available from most Bayesian programs such as MrBayes or BEAST. However, the approximation is extremely sensitive to the behavior of the MCMC sampler [52, 142]: if extremely low-likelihood values happen to be sampled from the posterior distribution, the harmonic mean will be dramatically affected. To date, a couple of more robust approximations have been described and were shown to be preferable to the harmonic mean estimator [52]. The first is based on thermodynamic integration [52] and is available in PhyloBayes (see Table 1). The second approximation [144] is based on a more direct computation [145], but its availability is currently limited to one specific model of evolution.

2.5. More realistic models

While model selection is fully justified on the ground of the bias-variance tradeoff, it should not be forgotten that all these models are simplified representations of the actual substitution process and are all therefore wrong. Stated differently, if AIC selects the GTR +Γ+I to analyze a data set, it should be clear that this conclusion does not imply that the data evolved under this model. All model selection procedures measure a relative model fit. One way to estimate adequacy or absolute model fit is to perform a parametric bootstrap test [146]: *first*, the selected model is compared with a multinomial model by means of a LRT whose test statistic is s (twice the log-likelihood difference); the following steps determine the distribution of s under the null hypothesis that the selected model was the generating model; *second*, the selected model is used to simulate a large number of data sets; *third*, the model selection procedure (LRT) is repeated on each simulated data set, and the corresponding test statistics s^* are recorded; *fourth*, the P-value is estimated as the number of times, the simulated s^* test statistics are more extreme ($>$, for a one-sided test) than the original value of s. The results of such tests suggest that the selected substitution model is generally not an adequate representation of the actual substitution process [85]. Of course, we do not need a model that incorporates all the minute biological features of evolutionary processes. As argued repeatedly (e.g., [147]), we need *useful* models that capture enough of reality of substitution processes to make accurate predictions and avoid systematic biases such as long-branch attraction [148].

More realistic models are obtained by accommodating heterogeneities in the evolutionary process at the level of both sites (space) and lineages (time). The simplest site-heterogeneous model is one, where the aligned data are partitioned, usually based on some prior information. For instance, first and second codon positions are known to evolve slower than third codon positions in protein-coding genes, or exposed residues might evolve faster than buried amino acids in globular proteins. A number of models were suggested to analyze such partitioned data sets (e.g., [149]); these models are implemented in most general-purpose software (e.g., PAML, PAUP, MrBayes) and can be combined with a "+Γ+I" component. A different approach consists in considering that sites can be binned in a number of rate categories; the use of a Dirichlet prior process then makes it possible both to determine the appropriate number of categories and to assign sites to these categories; the application of this method to protein-coding genes was able to recover the underlying codon structure of these genes [150]. However, several studies suggest that evolutionary patterns can be as heterogeneous within a priori partitions as among partitions [37, 151].

Lineage-heterogeneous models or heterotachous models [152] have attracted more attention. In one such approach, different models of evolution are assigned to the different branches of the tree [153], which can make these models extremely parameter-rich. Such a large number of parameters can potentially affect the accuracy of the phylogenetic inference (see the "bias-variance tradeoff" above) and present computational issues (long running times, large memory requirements, and convergence issues). Several simplifications can be made. One assumes that some sets of branches evolve under a particular process [153]. But now these branches must be assigned a priori, and both the determination of the number of sets and their placement on the tree can be difficult (but see Section 4 below for a solution to a similar question). At the other end of the spectrum of heterotachous models lies the simplest model known as the covarion model [154], where sites can either be variable along a branch, or not, and can switch between these two categories across time (e.g., [155], also described in a Bayesian framework [156]).

Between these two extremes are mixture models, which extend the covarion model by allowing more categories of sites. A number of formulations exist, where each site is assumed to have been generated by either several sets of branch lengths [157, 158] or by several rate matrices [37, 96, 151]. One particularity of these models is that they give a semi-parametric perspective to the phylogenetic estimation: if a single simple model cannot approximate a complex substitution process, the hope is that mixing several simple substitution models makes our models more realistic. In some applications, mixture models can also be used to avoid underestimating uncertainty, first when choosing a single model of evolution and then ignoring this uncertainty when estimating the phylogeny. The mixing therefore involves fitting at each site several sets of branch lengths, or several substitution models to the data, and combining these models using a certain weighting scheme. The difference between the numerous mixture models that have been described lies in the choice of the weight factors, and how these are obtained. In one approach, known as model averaging, the weights are determined a priori. A first possibility is to assume that all the models are equally probable, which does not work with an infinite number of models (individual weights are zero in this case). More critically in phylogenetics, this assumption is not coherent for nested models since larger models should be more likely than each submodel. A second possibility is to weight the models with respect to their probability of being the generating model given the data. For practical purposes, this posterior probability can be approximated by Akaike weights [96]. The difficulty here is that model averaging requires analyzing the data even for models that, a posteriori, turn out to have extremely small probabilities or

weights. This may be seen as a waste of resources (computing time and storage space).

2.6. Integrated Bayesian approaches

Mixture models can work within the framework of maximum likelihood, but the treatment of the weight factors is complicated. A sound alternative is to resort to a fully Bayesian approach. A prior distribution is set on the weight factors, and a special form of MCMC sampler whose Markov chain moves across models with different numbers of parameters, a reversible-jump MCMC sampler (RJ-MCMC), is constructed. The advantage of RJ-MCMC samplers is that they allow estimating the phylogeny while integrating over the uncertainty pertaining to the parameters of the substitution model and even integrating over the model itself [104]. Mixture models are available in BayesPhylogenies [37] for nucleotide models. Another Bayesian mixture model, named CAT for CATegories, was developed to analyze amino acid alignments. The CAT model recently proved successful in a number of empirical [159, 160] and simulation [161] studies in avoiding the artifact known as long-branch attraction [148]. This model is freely available in the PhyloBayes software (see Table 1).

All these models assume that each site evolve independently. The independence assumption greatly simplifies the computations, but is also highly unrealistic. Models that describe the evolution of doublets in RNA genes [162], triplets in codon models [163, 164], or other models with local or context dependencies [165–167] exist, but complete dependence models are still in their infancy and, so far, have only been implemented in a Bayesian framework [168, 169]. One particularly interesting feature of this approach is that complete dependence models incorporate information about the three-dimensional (3D) structure of proteins and therefore permit the explicit modeling of structural constraints or of any other site-interdependence pattern [170]. The incorporation of 3D structures also allows the establishment of a direct relationship between evolution at the DNA level and at the phenotypic level. This link between genotype and phenotype is established via a proxy that plays the role of a fitness function which, in retrospect, can be used to predict amino-acid sequences compatible with a given target structure, that is, to help in protein design [171].

3. DETECTING POSITIVE SELECTION

Fitness functions are however difficult to determine at the molecular level. In addition, while examples of adaptive evolution at the morphological level abound, from Darwin's finches in the Galapagos [172] to cichlid fishes in the East African lakes [173], the role of natural selection in shaping the evolution of genomes is much more controversial [147, 174]. First, the neutral theory of molecular evolution asserts that much of the variation at the DNA level is due to the random fixation of mutations with no selective advantage [175]. Second, a compelling body of evidence suggests that most of the genomic complexities have emerged by non-adaptive processes [176]. A number of statistical approaches exist either to test neutrality at the population level or to detect positive Darwinian evolution at the species level [147]. A shortcoming of neutrality tests is their dependence on a demographic model [177] and their sensitivity to processes of molecular evolution such as among-site rate variation [178]. They also do not model alternative hypotheses that would permit distinguishing negative selection from adaptive evolution. The development of demographic models based on Poisson random fields [179] and composite likelihoods [180] makes it possible both to estimate the strength of selection and to assess the impact of a variety of scenarios on allele frequency spectra [9]. But demographic singularities such as bottlenecks can still generate spurious signatures of positive selection [180, 181].

When effective population sizes are no longer a concern, for instance in studies at or above the species level, the detection of positive selection in protein-coding genes usually relies on codon models [163, 164] (see [182] for a review including methods based on amino-acid models). Codon models permit distinguishing between synonymous substitutions, which are likely to be neutral, and nonsynonymous substitutions, which are directly exposed to the action of selection. If synonymous and nonsynonymous substitutions accumulate at the same rate, then the protein-coding gene is likely to evolve neutrally. Alternatively, if nonsynonymous substitutions accumulate slower than synonymous substitutions, it must be because nonsynonymous substitutions are deleterious and this suggests the action of purifying selection. Conversely, the accumulation of nonsynonymous substitutions faster than synonymous substitutions suggests the action of positive selection. The nonsynonymous to synonymous rate ratio, denoted $\omega = d_N/d_S$, is therefore interpreted as a measure of selection at the protein level, with $\omega = 1$, <1 and >1 indicating neutral evolution, negative or positive selection, respectively. This ratio is also denoted K_a/K_s, in particular in studies that rely on counts of nonsynonymous and synonymous sites (e.g., [183]). An extension exists to detect selection in noncoding regions [184], and a promising phylogenetic hidden Markov or phylo-HMM model permits detection of selection in overlapping genes [185].

These rate ratios can be estimated by a number of methods implemented in MEGA, DAMBE, HyPhy [42], and PAML. The most intuitive methods, called counting methods, work in three steps: (i) count synonymous and nonsynonymous sites, (ii) count the observed differences at these sites, and (iii) apply corrections for multiple substitutions [186]. Counting methods are however not optimal in the sense that most work on pairs of sequences and therefore, just like neighbor-joining, fail to account for all the information contained in an alignment. In addition, simulations suggest that counting methods can be sensitive to a variety of biases such as unequal transition and transversion rates, or uneven base, or codon frequencies [187]. Counting methods that incorporate these biases perform generally better than those that do not, but the maximum likelihood method still appears more robust to sever biases [187]. In addition, the maximum likelihood method that accounts for all the information in a data set has good power and good accuracy to detect positive selection [188, 189].

However, the first studies using these methods found little evidence for adaptive evolution essentially because they were averaging ω rate ratios over both lineages and sites [147]. Branch models were then developed [190, 191] quickly followed by site models [192–196] and by branch-site models [189, 197]. All these approaches, as implemented in PAML, rely on likelihood ratio tests to detect adaptive evolution: a model where adaptive evolution is permitted is compared with a null model where ω cannot be greater than one. Simulations show that some of these tests are conservative [189], so that detection of adaptive evolution should be safe as long as convergence of the analyses is carefully checked [198], including in large-scale analyses [199]. If the model allowing adaptive evolution explains the data significantly better than the null model, then an empirical Bayes approach can be used to identify which sites are likely to evolve adaptively [192]. The empirical Bayes approach relies on estimates of the model parameters, which can have large sampling errors in small data sets. Because these sampling errors can cause the empirical Bayes site identification to be unreliable [200], a Bayes empirical Bayes approach was proposed and was shown to have good power and low-false positive rates [201]. Full Bayesian approaches that allow for uncertain parameter estimates were also proposed [202]. Yet, simulations showed that they did not improve further on Bayes empirical Bayes estimates [203], so that the computational overhead incurred by full Bayes methods may not be necessary in this case. One particular case, where a Bayesian approach is however required, is to tell the signature of adaptive evolution from that of recombination, as these two processes can leave similar signals in DNA sequences. Indeed, simulations show that recombination can lead to false positive rates as large as 90% when trying to detect adaptive evolution [204]. The codon model with recombination implemented in OmegaMap [48] can then be used to tease apart these two processes (e.g., see [205]).

4. ESTIMATING DIVERGENCE TIMES BETWEEN SPECIES

The estimation of the dates when species diverged is often perceived to be as important as estimating the phylogeny itself. This explains why so-called "dating methods" were first wished for when molecular phylogenies were first reconstructed [206]. In spite of over four decades of history, molecular dating has only recently seen new developments. One of the reasons for this slow progress is that, unlike the other parts of phylogenetic analysis, divergence times are parameters that cannot be estimated directly. Only sitewise likelihood values and distances between pairs of sequences are identifiable, that is, directly estimable. Distances are expressed as a number of substitutions per site (sub/site) and can be decomposed as the product of two quantities: a rate of evolution (sub/site/unit of time) and a time duration (unit of time). As a result, time durations and, likewise, divergence times cannot be estimated without making an additional assumption on the rates of evolution. The simplest assumption is to posit that rates are constant in time, which is known as the molecular clock hypothesis [207]. This hypothesis can be tested, for instance, with PAUP or PAML, by means of a likelihood ratio test that compares a constrained model (clock) with an unconstrained model (no clock). These two models are nested, so that twice the log-likelihood difference asymptotically follows a χ^2 distribution. If n sequences are analyzed, the constrained model estimates $n - 1$ divergence times, while the unconstrained model estimates $2n - 3$ branch lengths. The degree of freedom of this test is then $(2n - 3) - (n - 1) = n - 2$ [4]. The systematic test of the molecular clock assumption on recent data shows that this hypothesis is too often untenable [208].

The most recent work has then focused on relaxing this assumption, and three different directions have emerged [209]. A first possibility is to relax the clock *globally* on the phylogeny, but to assume that the hypothesis still holds *locally* for closely related species [210–212]. Recent developments of these local clock models now allow the use of multiple calibration points and of multiple genes [213], the automatic placement of the clocks on the tree [214] and the estimation of the number of local clocks [209]. PAML can be used for most of these computations. However, local clock models still tend to underestimate rapid rate change [209]. The second possibility to relax the global clock assumption is to assume that rates of evolution evolve in an autocorrelated manner along lineages and to minimize the amount of rate change over the entire phylogeny. The most popular approach in the plant community is Sanderson's penalized likelihood [215], implemented in r8s [55]. This approach performs well on data sets for which the actual fossil dates are known [216] but still tends to underestimate the actual amount of rate change [209].

Bayesian methods appear today as the emerging approach to estimate divergence times. Taking inspiration from Sanderson's pioneering work [217], Thorne et al. developed a Bayesian framework where rates of evolution change in an autocorrelated manner across lineages [45–47]: the rate of evolution of a branch depends on the rate of evolution of its parental branch; the branches emanating from the root require a special treatment. These Bayesian models work by modeling how rates of evolution change in time (rate prior), and how the speciation/population process shapes the distribution of divergence times (speciation prior). These prior distributions can actually be interpreted as penalty functions [45, 209], and they can have simple or more complicated forms [218]. The Multidivtime program [45–47] is extremely quick to analyze data thanks to the use of a multivariate normal approximation of the likelihood surface. It assumes that rates of evolution change following a stationary lognormal prior distribution. Further work suggested that it might not always be the best performing rate prior [218–220], but these latter studies had two potential shortcomings: (i) they were based on a speciation prior that was so strong that it biased divergence times towards the age of the fossil root [219, 221], and (ii) they used a statistical procedure, the posterior Bayes factor [222], that is potentially inconsistent. One potential limitation of the Bayesian approach described so far is its dependence on one single tree topology, which must be either known ahead of time or estimated by other means. Recently, Drummond et al. found a way to relax this requirement by

positing that rates of evolution are uncorrelated across lineages, while all the branches of the tree are constrained to follow exactly the same rate prior [223]. As a result, their approach is able to estimate the most probable tree (given the data and the substitution model), the divergence times and the position of the root even without any outgroup or without resorting to a nonreversible model of substitution [224]. Drummond et al. further argue that the use of explicit models of rate variation over time might contribute to improved phylogenetic inference [223]. In addition, when the focus is on estimating divergence times, a recent analysis suggests that this uncorrelated model of rate change could outperform the methods described above to accommodate rapid rate change among lineages [209]. Implemented in BEAST, this approach offers a variety of substitution models and prior distributions and presents a graphic user interface that will appeal to numerous researchers [39].

5. CHALLENGES AND PERSPECTIVES

With the advent of high-throughput sequencing technologies such as the whole-genome shotgun approach by pyrosequencing [225], fast, cheap, and accurate genomic information is becoming available for a growing number of species [226]. If low coverage limits the complete assembly of many genome projects, it still allows the quick access to draft genomes for a growing number of species [227]. As a result, phylogenetic inference can now incorporate large numbers of expressed sequence tags (ESTs), genes [228], and occasionally complete genomes [229]. The motivation for developing these so-called phylogenomic approaches is their presumed ability to return fully resolved and well-supported trees by decreasing both sampling errors [230] and misleading signals due for instance to horizontal gene transfer [231] or to hidden paralogy [232]. In practice, these large-scale studies can give the impression that incongruence is resolved [228], but they also can fail to address systematic errors due to the use of too simple models [233]. If the genes incorporated in phylogenomic studies are often concatenated to limit the number of parameters entering the model, it remains important to allow sitewise heterogeneities [234]. If partition models can reduce systematic biases [234], Bayesian mixture models such as CAT [151] appear to be robust to long-branch attraction [159], a rampant issue in phylogenomics [235]. All together, the accumulation of genomic data and these latest methodological developments seem to make the reconstruction of the tree of life finally within reach. In comparison, dating the tree of life is still in its infancy, even if a number of initiatives such as the TimeTree server are being developed [236]. These resources are limited to some vertebrates but will hopefully soon be extended to include other large taxonomic groups such as plants. To achieve this goal, however, phylogenetic studies should systematically incorporate divergence times, as is now routine in some research communities (e.g., [237]). This joint estimation of time and trees is today facilitated by the availability of user-friendly programs such as BEAST. The near future will probably see the development of mixture models for molecular dating and more sophisticated models that integrate most of the topics discussed here from sequence alignment to detection of sites under selection into one single but yet user-friendly [238] toolbox.

ACKNOWLEDGMENTS

Jeff Thorne provided insightful comments and suggestions, and two anonymous reviewers helped in improving the original manuscript. Support was provided by the Natural Sciences Research Council of Canada (DG-311625 to SAB and DG-261252 to XX).

REFERENCES

[1] C. Darwin, *On the Origin of Species by Means of Natural Selection*, J. Murray, London, UK, 1859.

[2] R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy*, W. H. Freeman, San Francisco, Calif, USA, 1963.

[3] L. L. Cavalli-Sforza, I. Barrai, and A. W. Edwards, "Analysis of human evolution under random genetic drift," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 29, pp. 9–20, 1964.

[4] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, Mass, USA, 2004.

[5] H. Glenner, A. J. Hansen, M. V. Sørensen, F. Ronquist, J. P. Huelsenbeck, and E. Willerslev, "Bayesian inference of the metazoan phylogeny: a combined molecular and morphological approach," *Current Biology*, vol. 14, no. 18, pp. 1644–1649, 2004.

[6] B. E. Pfeil, J. A. Schlueter, R. C. Shoemaker, and J. J. Doyle, "Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families," *Systematic Biology*, vol. 54, no. 3, pp. 441–454, 2005.

[7] E. R. Chare and E. C. Holmes, "A phylogenetic survey of recombination frequency in plant RNA viruses," *Archives of Virology*, vol. 151, no. 5, pp. 933–946, 2006.

[8] H. Philippe and C. J. Douady, "Horizontal gene transfer and phylogenetics," *Current Opinion in Microbiology*, vol. 6, no. 5, pp. 498–505, 2003.

[9] R. Nielsen, C. Bustamante, A. G. Clark, et al., "A scan for positively selected genes in the genomes of humans and chimpanzees," *PLoS Biology*, vol. 3, no. 6, p. e170, 2005.

[10] S. R. Ramírez, B. Gravendeel, R. B. Singer, C. R. Marshall, and N. E. Pierce, "Dating the origin of the Orchidaceae from a fossil orchid with its pollinator," *Nature*, vol. 448, no. 7157, pp. 1042–1045, 2007.

[11] R. D. Knight, S. J. Freeland, and L. F. Landweber, "Rewiring the keyboard: evolvability of the genetic code," *Nature Reviews Genetics*, vol. 2, no. 1, pp. 49–58, 2001.

[12] J. Antonovics, M. E. Hood, and C. H. Baker, "Molecular virology: was the 1918 flu avian in origin?" *Nature*, vol. 440, no. 7088, p. E9, 2006, discussion E9-10.

[13] A. P. Jackson and M. A. Charleston, "A cophylogenetic perspective of RNA-virus evolution," *Molecular Biology and Evolution*, vol. 21, no. 1, pp. 45–57, 2004.

[14] J. P. Huelsenbeck, B. Rannala, and B. Larget, "A Bayesian framework for the analysis of cospeciation," *Evolution*, vol. 54, no. 2, pp. 352–364, 2000.

[15] M. Hajibabaei, G. A. C. Singer, P. D. N. Hebert, and D. A. Hickey, "DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics," *Trends in Genetics*, vol. 23, no. 4, pp. 167–172, 2007.

[16] S.-J. Luo, J.-H. Kim, W. E. Johnson, et al., "Phylogeography and genetic ancestry of tigers (*Panthera tigris*)," *PLoS Biology*, vol. 2, no. 12, p. e442, 2004.

[17] C. J. Howe, A. C. Barbrook, M. Spencer, P. Robinson, B. Bordalejo, and L. R. Mooney, "Manuscript evolution," *Endeavour*, vol. 25, no. 3, pp. 121–126, 2001.

[18] R. D. Gray and Q. D. Atkinson, "Language-tree divergence times support the Anatolian theory of Indo-European origin," *Nature*, vol. 426, no. 6965, pp. 435–439, 2003.

[19] D. M. Hillis and J. P. Huelsenbeck, "Support for dental HIV transmission," *Nature*, vol. 369, no. 6475, pp. 24–25, 1994.

[20] A. Salas, H.-J. Bandelt, V. Macaulay, and M. B. Richards, "Phylogeographic investigations: the role of trees in forensic genetics," *Forensic Science International*, vol. 168, no. 1, pp. 1–13, 2007.

[21] D. Sankoff and J. H. Nadeau, "Chromosome rearrangements in evolution: from gene order to genome sequence and back," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 20, pp. 11188–11189, 2003.

[22] D. L. Swofford, P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers, "Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods," *Systematic Biology*, vol. 50, no. 4, pp. 525–539, 2001.

[23] M. Holder and P. O. Lewis, "Phylogeny estimation: traditional and Bayesian approaches," *Nature Reviews Genetics*, vol. 4, no. 4, pp. 275–284, 2003.

[24] A. Siepel and D. Haussler, "Phylogenetic hidden Markov models," in *Statistical Methods in Molecular Evolution*, R. Nielsen, Ed., pp. 325–351, Springer, New York, NY, USA, 2005.

[25] M. Pagel and A. Meade, "Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo," *American Naturalist*, vol. 167, no. 6, pp. 808–825, 2006.

[26] S. Kumar and A. Filipski, "Multiple sequence alignment: in pursuit of homologous DNA positions," *Genome Research*, vol. 17, no. 2, pp. 127–135, 2007.

[27] C. Notredame, "Recent evolutions of multiple sequence alignment algorithms," *PLoS Computational Biology*, vol. 3, no. 8, p. e123, 2007.

[28] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 368–373, 2006.

[29] X. Xia and Z. Xie, "DAMBE: software package for data analysis in molecular biology and evolution," *Journal of Heredity*, vol. 92, no. 4, pp. 371–373, 2001.

[30] S. Kumar, K. Tamura, and M. Nei, "MEGA: molecular evolutionary genetics analysis software for microcomputers," *Computer Applications in the Biosciences*, vol. 10, no. 2, pp. 189–191, 1994.

[31] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[32] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: probabilistic consistency-based multiple sequence alignment," *Genome Research*, vol. 15, no. 2, pp. 330–340, 2005.

[33] I. M. Wallace, G. Blackshields, and D. G. Higgins, "Multiple sequence alignments," *Current Opinion in Structural Biology*, vol. 15, no. 3, pp. 261–266, 2005.

[34] I. M. Wallace, O. O'Sullivan, D. G. Higgins, and C. Notredame, "M-Coffee: combining multiple sequence alignment methods with T-Coffee," *Nucleic Acids Research*, vol. 34, no. 6, pp. 1692–1699, 2006.

[35] B. G. Hall, *Phylogenetic Trees Made Easy: A How-to Manual*, Sinauer Associates, Sunderland, Mass, USA, 2008.

[36] B. Larget and D. L. Simon, "Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees," *Molecular Biology and Evolution*, vol. 16, no. 6, pp. 750–759, 1999.

[37] M. Pagel and A. Meade, "A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data," *Systematic Biology*, vol. 53, no. 4, pp. 571–581, 2004.

[38] B. D. Redelings and M. A. Suchard, "Joint Bayesian estimation of alignment and phylogeny," *Systematic Biology*, vol. 54, no. 3, pp. 401–418, 2005.

[39] A. J. Drummond and A. Rambaut, "BEAST: Bayesian evolutionary analysis by sampling trees," *BMC Evolutionary Biology*, vol. 7, article 214, pp. 1–8, 2007.

[40] H. Shimodaira and M. Hasegawa, "CONSEL: for assessing the confidence of phylogenetic tree selection," *Bioinformatics*, vol. 17, no. 12, pp. 1246–1247, 2001.

[41] D. Zwickl, "Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion," Ph.D. thesis, University of Texas at Austin, Austin, Tex, USA, 2006.

[42] S. L. Kosakovsky Pond, S. D. W. Frost, and S. V. Muse, "HyPhy: hypothesis testing using phylogenies," *Bioinformatics*, vol. 21, no. 5, pp. 676–679, 2005.

[43] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: Bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[44] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist, "Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference," *Bioinformatics*, vol. 20, no. 3, pp. 407–415, 2004.

[45] J. L. Thorne, H. Kishino, and I. S. Painter, "Estimating the rate of evolution of the rate of molecular evolution," *Molecular Biology and Evolution*, vol. 15, no. 12, pp. 1647–1657, 1998.

[46] H. Kishino, J. L. Thorne, and W. J. Bruno, "Performance of a divergence time estimation method under a probabilistic model of rate evolution," *Molecular Biology and Evolution*, vol. 18, no. 3, pp. 352–361, 2001.

[47] J. L. Thorne and H. Kishino, "Divergence time and evolutionary rate estimation with multilocus data," *Systematic Biology*, vol. 51, no. 5, pp. 689–702, 2002.

[48] D. J. Wilson and G. McVean, "Estimating diversifying selection and functional constraint in the presence of recombination," *Genetics*, vol. 172, no. 3, pp. 1411–1425, 2006.

[49] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.

[50] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[51] D. L. Swofford, *PAUP*: Phylogenetic Analysis Using Parsimony (and other Methods) 4.0 Beta*, Sinauer Associates, Sunderland, Mass, USA, 10th edition, 2002.

[52] N. Lartillot and H. Philippe, "Computing Bayes factors using thermodynamic integration," *Systematic Biology*, vol. 55, no. 2, pp. 195–207, 2006.

[53] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[54] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 2006.

[55] M. J. Sanderson, "r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock," *Bioinformatics*, vol. 19, no. 2, pp. 301–302, 2003.

[56] K. M. Kjer, "Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs," *Molecular Phylogenetics and Evolution*, vol. 4, no. 3, pp. 314–330, 1995.

[57] C. Notredame, E. A. O'Brien, and D. G. Higgins, "RAGA: RNA sequence alignment by genetic algorithm," *Nucleic Acids Research*, vol. 25, no. 22, pp. 4570–4580, 1997.

[58] R. E. Hickson, C. Simon, and S. W. Perrey, "The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence," *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 530–539, 2000.

[59] X. Xia, "Phylogenetic relationship among horseshoe crab species: effect of substitution models on phylogenetic analyses," *Systematic Biology*, vol. 49, no. 1, pp. 87–100, 2000.

[60] X. Xia, Z. Xie, and K. M. Kjer, "18S ribosomal RNA and tetrapod phylogeny," *Systematic Biology*, vol. 52, no. 3, pp. 283–295, 2003.

[61] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[62] M. A. Larkin, G. Blackshields, N. P. Brown, et al., "Clustal W and clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[63] T. Golubchik, M. J. Wise, S. Easteal, and L. S. Jermiin, "Mind the gaps: evidence of bias in estimates of multiple sequence alignments," *Molecular Biology and Evolution*, vol. 24, no. 11, pp. 2433–2442, 2007.

[64] G. Landan and D. Graur, "Heads or tails: a simple reliability check for multiple sequence alignments," *Molecular Biology and Evolution*, vol. 24, no. 6, pp. 1380–1383, 2007.

[65] A. S. Schwartz, E. W. Myers, and L. Pachter, "Alignment metric accuracy," http://arxiv.org/abs/q-bio.QM/0510052, 2005.

[66] J. Zhu, J. S. Liu, and C. E. Lawrence, "Bayesian adaptive sequence alignment algorithms," *Bioinformatics*, vol. 14, no. 1, pp. 25–39, 1998.

[67] I. Holmes and W. J. Bruno, "Evolutionary HMMs: a Bayesian approach to multiple alignment," *Bioinformatics*, vol. 17, no. 9, pp. 803–820, 2001.

[68] J. L. Jensen and J. Hein, "Gibbs sampler for statistical multiple alignment," *Statistica Sinica*, vol. 15, no. 4, pp. 889–907, 2005.

[69] M. Clamp, J. Cuff, S. M. Searle, and G. J. Barton, "The Jalview Java alignment editor," *Bioinformatics*, vol. 20, no. 3, pp. 426–427, 2004.

[70] D. Sankoff and R. Cedergren, "Simultaneous comparison of three or more sequences related by a tree," in *Time Wraps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff and R. Cedergren, Eds., pp. 253–264, Addison-Wesley, Reading, Mass, USA, 1983.

[71] J. Hein, "A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given," *Molecular Biology and Evolution*, vol. 6, no. 6, pp. 649–668, 1989.

[72] G. Lunter, I. Miklós, A. Drummond, J. L. Jensen, and J. Hein, "Bayesian coestimation of phylogeny and sequence alignment," *BMC Bioinformatics*, vol. 6, article 83, pp. 1–10, 2005.

[73] K. M. Wong, M. A. Suchard, and J. P. Huelsenbeck, "Alignment uncertainty and genomic analysis," *Science*, vol. 319, no. 5862, pp. 473–476, 2008.

[74] G. Lunter, A. Rocco, N. Mimouni, A. Heger, A. Caldeira, and J. Hein, "Uncertainty in homology inferences: assessing and improving genomic sequence alignment," *Genome Research*, vol. 18, no. 2, pp. 298–309, 2008.

[75] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, NY, USA, 2000.

[76] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," in *Mammalian Protein Metabolism*, H. N. Munro, Ed., pp. 21–121, Academic Press, New York, NY, USA, 1969.

[77] M. Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," *Journal of Molecular Evolution*, vol. 16, no. 2, pp. 111–120, 1980.

[78] M. Hasegawa, H. Kishino, and T. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution*, vol. 22, no. 2, pp. 160–174, 1985.

[79] K. Tamura and M. Nei, "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees," *Molecular Biology and Evolution*, vol. 10, no. 3, pp. 512–526, 1993.

[80] S. Tavare, "Some probabilistic and statistical problems on the analysis of DNA sequences," in *Lectures on Mathematics in the Life Sciences*, vol. 17, pp. 57–86, American Mathematical Society, Providence, RI, USA, 1986.

[81] Z. Yang, "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites," *Molecular Biology and Evolution*, vol. 10, no. 6, pp. 1396–1401, 1993.

[82] Z. Yang, "Estimating the pattern of nucleotide substitution," *Journal of Molecular Evolution*, vol. 39, no. 1, pp. 105–111, 1994.

[83] N. Goldman and S. Whelan, "A novel use of equilibrium frequencies in models of sequence evolution," *Molecular Biology and Evolution*, vol. 19, no. 11, pp. 1821–1831, 2002.

[84] P. Liò and N. Goldman, "Models of molecular evolution and phylogeny," *Genome Research*, vol. 8, no. 12, pp. 1233–1244, 1998.

[85] S. Whelan, P. Liò, and N. Goldman, "Molecular phylogenetics: state-of-the-art methods for looking into the past," *Trends in Genetics*, vol. 17, no. 5, pp. 262–272, 2001.

[86] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach*, Springer, New York, NY, USA, 2002.

[87] W. J. Bruno and A. L. Halpern, "Topological bias and inconsistency of maximum likelihood using wrong models," *Molecular Biology and Evolution*, vol. 16, no. 4, pp. 564–566, 1999.

[88] D. Posada and K. A. Crandall, "Selecting the best-fit model of nucleotide substitution," *Systematic Biology*, vol. 50, no. 4, pp. 580–601, 2001.

[89] D. R. Cox, "Further results on tests of separate families of hypotheses," *Journal of the Royal Statistical Society. Series B*, vol. 24, no. 2, pp. 406–424, 1962.

[90] N. Goldman and S. Whelan, "Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics," *Molecular Biology and Evolution*, vol. 17, no. 6, pp. 975–978, 2000.

[91] M. Anisimova and O. Gascuel, "Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative," *Systematic Biology*, vol. 55, no. 4, pp. 539–552, 2006.

[92] D. Posada and K. A. Crandall, "MODELTEST: testing the model of DNA substitution," *Bioinformatics*, vol. 14, no. 9, pp. 817–818, 1998.

[93] E. Paradis, J. Claude, and K. Strimmer, "APE: analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289–290, 2004.

[94] D. Posada, "ModelTest server: a web-based tool for the statistical selection of models of nucleotide substitution online," *Nucleic Acids Research*, vol. 34, web server issue, pp. W700–W703, 2006.

[95] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[96] D. Posada and T. R. Buckley, "Model selection and model averaging in phylogenetics: advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests," *Systematic Biology*, vol. 53, no. 5, pp. 793–808, 2004.

[97] D. Pol, "Empirical problems of the hierarchical likelihood ratio test for model selection," *Systematic Biology*, vol. 53, no. 6, pp. 949–962, 2004.

[98] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[99] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[100] V. N. Minin, Z. Abdo, P. Joyce, and J. Sullivan, "Performance-based selection of likelihood models for phylogeny estimation," *Systematic Biology*, vol. 52, no. 5, pp. 674–683, 2003.

[101] Z. Abdo, V. N. Minin, P. Joyce, and J. Sullivan, "Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation," *Molecular Biology and Evolution*, vol. 22, no. 3, pp. 691–703, 2005.

[102] L. Bao, H. Gu, K. A. Dunn, and J. P. Bielawski, "Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on their application to gene and genome data," *BMC Evolutionary Biology*, vol. 7, supplement 1, p. S5, 2007.

[103] M. A. Suchard, R. E. Weiss, and J. S. Sinsheimer, "Bayesian selection of continuous-time Markov chain evolutionary models," *Molecular Biology and Evolution*, vol. 18, no. 6, pp. 1001–1013, 2001.

[104] J. P. Huelsenbeck, B. Larget, and M. E. Alfaro, "Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 1123–1133, 2004.

[105] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[106] O. Gascuel and M. Steel, "Neighbor-joining revealed," *Molecular Biology and Evolution*, vol. 23, no. 11, pp. 1997–2000, 2006.

[107] W. J. Bruno, N. D. Socci, and A. L. Halpern, "Weighted neighbor-joining: a likelihood-based approach to distance-based phylogeny reconstruction," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 189–197, 2000.

[108] S. L. Baldauf, "Phylogeny for the faint of heart: a tutorial," *Trends in Genetics*, vol. 19, no. 6, pp. 345–351, 2003.

[109] L. L. Cavalli-Sforza and A. W. F. Edwards, "Phylogenetic analysis. Models and estimation procedures," *American Journal of Human Genetics*, vol. 19, no. 3, part 1, pp. 233–257, 1967.

[110] S. Whelan, "New approaches to phylogenetic tree search and their application to large numbers of protein alignments," *Systematic Biology*, vol. 56, no. 5, pp. 727–740, 2007.

[111] M. T. Holder, P. O. Lewis, D. L. Swofford, and B. Larget, "Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics," *Systematic Biology*, vol. 54, no. 6, pp. 961–965, 2005.

[112] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[113] D. M. Hillis and J. J. Bull, "An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis," *Systematic Biology*, vol. 42, no. 2, pp. 182–192, 1993.

[114] J. Felsenstein and H. Kishino, "Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull," *Systematic Biology*, vol. 42, no. 2, pp. 193–200, 1993.

[115] Z. Yang and B. Rannala, "Branch-length prior influences Bayesian posterior probability of phylogeny," *Systematic Biology*, vol. 54, no. 3, pp. 455–470, 2005.

[116] V. Berry and O. Gascuel, "On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain," *Molecular Biology and Evolution*, vol. 13, no. 7, pp. 999–1011, 1996.

[117] B. Efron, E. Halloran, and S. Holmes, "Bootstrap confidence levels for phylogenetic trees," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 14, pp. 7085–7090, 1996.

[118] B. Mau, M. A. Newton, and B. Larget, "Bayesian phylogenetic inference via Markov chain Monte Carlo methods," *Biometrics*, vol. 55, no. 1, pp. 1–12, 1999.

[119] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback, "Bayesian inference of phylogeny and its impact on evolutionary biology," *Science*, vol. 294, no. 5550, pp. 2310–2314, 2001.

[120] W. J. Murphy, E. Eizirik, S. J. O'Brien, et al., "Resolution of the early placental mammal radiation using Bayesian phylogenetics," *Science*, vol. 294, no. 5550, pp. 2348–2351, 2001.

[121] C. J. Douady, F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery, "Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability," *Molecular Biology and Evolution*, vol. 20, no. 2, pp. 248–254, 2003.

[122] M. P. Cummings, S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka, "Comparing bootstrap and posterior probability values in the four-taxon case," *Systematic Biology*, vol. 52, no. 4, pp. 477–487, 2003.

[123] P. Erixon, B. Svennblad, T. Britton, and B. Oxelman, "Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics," *Systematic Biology*, vol. 52, no. 5, pp. 665–673, 2003.

[124] B. Svennblad, P. Erixon, B. Oxelman, and T. Britton, "Fundamental differences between the methods of maximum likelihood and maximum posterior probability in phylogenetics," *Systematic Biology*, vol. 55, no. 1, pp. 116–121, 2006.

[125] J. P. Huelsenbeck and B. Rannala, "Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models," *Systematic Biology*, vol. 53, no. 6, pp. 904–913, 2004.

[126] P. O. Lewis, M. T. Holder, and K. E. Holsinger, "Polytomies and Bayesian phylogenetic inference," *Systematic Biology*, vol. 54, no. 2, pp. 241–253, 2005.

[127] B. Kolaczkowski and J. W. Thornton, "Effects of branch length uncertainty on Bayesian posterior probabilities for phylogenetic hypotheses," *Molecular Biology and Evolution*, vol. 24, no. 9, pp. 2108–2118, 2007.

[128] M. Steel and F. A. Matsen, "The Bayesian "star paradox" persists for long finite sequences," *Molecular Biology and Evolution*, vol. 24, no. 4, pp. 1075–1079, 2007.

[129] Z. Yang, "Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1639–1655, 2007.

[130] B. Kolaczkowski and J. W. Thornton, "Is there a star tree paradox?" *Molecular Biology and Evolution*, vol. 23, no. 10, pp. 1819–1823, 2006.

[131] E. Mossel and E. Vigoda, "Phylogenetic MCMC algorithms are misleading on mixtures of trees," *Science*, vol. 309, no. 5744, pp. 2207–2209, 2005.

[132] F. Ronquist, B. Larget, J. P. Huelsenbeck, J. B. Kadane, D. Simon, and P. van der Mark, "Comment on "Phylogenetic MCMC algorithms are misleading on mixtures of trees"," *Science*, vol. 312, no. 5772, p. 367, 2006.

[133] W. C. Wheeler and K. M. Pickett, "Topology-Bayes versus clade-Bayes in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 25, no. 2, pp. 447–453, 2008.

[134] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees: hardness and approximation," *Bioinformatics*, vol. 21, supplement 1, pp. i97–i106, 2005.

[135] M. J. Donoghue, "Progress and prospects in reconstructing plant phylogeny," *Annals of the Missouri Botanical Garden*, vol. 81, no. 3, pp. 405–418, 1994.

[136] S. Aris-Brosou, "Least and most powerful phylogenetic tests to elucidate the origin of the seed plants in the presence of conflicting signals under misspecified models," *Systematic Biology*, vol. 52, no. 6, pp. 781–793, 2003.

[137] H. Kishino and M. Hasegawa, "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoide," *Journal of Molecular Evolution*, vol. 29, no. 2, pp. 170–179, 1989.

[138] N. Goldman, J. P. Anderson, and A. G. Rodrigo, "Likelihood-based tests of topologies in phylogenetics," *Systematic Biology*, vol. 49, no. 4, pp. 652–670, 2000.

[139] H. Shimodaira and M. Hasegawa, "Multiple comparisons of log-likelihoods with applications to phylogenetic inference," *Molecular Biology and Evolution*, vol. 16, no. 8, pp. 1114–1116, 1999.

[140] H. Shimodaira, "An approximately unbiased test of phylogenetic tree selection," *Systematic Biology*, vol. 51, no. 3, pp. 492–508, 2002.

[141] S. Aris-Brosou, "How Bayes tests of molecular phylogenies compare with frequentist approaches," *Bioinformatics*, vol. 19, no. 5, pp. 618–624, 2003.

[142] A. E. Raftery, "Hypothesis testing and model selection," in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds., pp. 163–187, Chapman & Hall, Boca Raton, Fla, USA, 1996.

[143] J. A. A. Nylander, F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey, "Bayesian phylogenetic analysis of combined data," *Systematic Biology*, vol. 53, no. 1, pp. 47–67, 2004.

[144] S. C. Choi, A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne, "Quantifying the impact of protein tertiary structure on molecular evolution," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1769–1782, 2007.

[145] S. Chib and I. Jeliazkov, "Marginal likelihood from the Metropolis-Hastings output," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 270–281, 2001.

[146] N. Goldman, "Statistical tests of models of DNA substitution," *Journal of Molecular Evolution*, vol. 36, no. 2, pp. 182–198, 1993.

[147] Z. Yang, *Computational Molecular Evolution*, Oxford University Press, Oxford, UK, 2006.

[148] J. Felsenstein, "Cases in which parsimony or compatibility methods will be positively misleading," *Systematic Zoology*, vol. 27, no. 4, pp. 401–410, 1978.

[149] Z. Yang, "Maximum-likelihood models for combined analyses of multiple sequence data," *Journal of Molecular Evolution*, vol. 42, no. 5, pp. 587–596, 1996.

[150] J. P. Huelsenbeck and M. A. Suchard, "A nonparametric method for accommodating and testing across-site rate variation," *Systematic Biology*, vol. 56, no. 6, pp. 975–987, 2007.

[151] N. Lartillot and H. Philippe, "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 1095–1109, 2004.

[152] P. Lopez, D. Casane, and H. Philippe, "Heterotachy, an important process of protein evolution," *Molecular Biology and Evolution*, vol. 19, no. 1, pp. 1–7, 2002.

[153] Z. Yang and D. Roberts, "On the use of nucleic acid sequences to infer early branchings in the tree of life," *Molecular Biology and Evolution*, vol. 12, no. 3, pp. 451–458, 1995.

[154] W. M. Fitch and E. Markowitz, "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution," *Biochemical Genetics*, vol. 4, no. 5, pp. 579–593, 1970.

[155] C. Tuffley and M. Steel, "Modeling the covarion hypothesis of nucleotide substitution," *Mathematical Biosciences*, vol. 147, no. 1, pp. 63–91, 1998.

[156] J. P. Huelsenbeck, "Testing a covariotide model of DNA substitution," *Molecular Biology and Evolution*, vol. 19, no. 5, pp. 698–707, 2002.

[157] B. Kolaczkowski and J. W. Thornton, "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogenous," *Nature*, vol. 431, no. 7011, pp. 980–984, 2004.

[158] M. Spencer, E. Susko, and A. J. Roger, "Likelihood, parsimony, and heterogeneous evolution," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1161–1164, 2005.

[159] N. Lartillot, H. Brinkmann, and H. Philippe, "Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model," *BMC Evolutionary Biology*, vol. 7, supplement 1, p. S4, 2007.

[160] E. Jiménez-Guri, H. Philippe, B. Okamura, and P. W. H. Holland, " *Buddenbrockia* is a cnidarian worm," *Science*, vol. 317, no. 5834, pp. 116–118, 2007.

[161] H. Philippe, Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc, "Heterotachy and long-branch attraction in phylogenetics," *BMC Evolutionary Biology*, vol. 5, article 50, pp. 1–8, 2005.

[162] M. Schöniger and A. Von Haeseler, "A stochastic model for the evolution of autocorrelated DNA sequences," *Molecular Phylogenetics and Evolution*, vol. 3, no. 3, pp. 240–247, 1994.

[163] S. V. Muse and B. S. Gaut, "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome," *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 715–724, 1994.

[164] N. Goldman and Z. Yang, "A codon-based model of nucleotide substitution for protein-coding DNA sequences," *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 725–736, 1994.

[165] A. Siepel and D. Haussler, "Phylogenetic estimation of context-dependent substitution rates by maximum likelihood," *Molecular Biology and Evolution*, vol. 21, no. 3, pp. 468–488, 2004.

[166] D. G. Hwang and P. Green, "Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 39, pp. 13994–14001, 2004.

[167] O. F. Christensen, A. Hobolth, and J. L. Jensen, "Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates," *Journal of Computational Biology*, vol. 12, no. 9, pp. 1166–1182, 2005.

[168] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne, "Protein evolution with dependence among codons due to tertiary structure," *Molecular Biology and Evolution*, vol. 20, no. 10, pp. 1692–1704, 2003.

[169] N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe, "Site interdependence attributed to tertiary structure in amino acid sequence evolution," *Gene*, vol. 347, no. 2, pp. 207–217, 2005.

[170] N. Rodrigue, H. Philippe, and N. Lartillot, "Assessing site-interdependent phylogenetic models of sequence evolution," *Molecular Biology and Evolution*, vol. 23, no. 9, pp. 1762–1775, 2006.

[171] C. L. Kleinman, N. Rodrigue, C. Bonnard, H. Philippe, and N. Lartillot, "A maximum likelihood framework for protein design," *BMC Bioinformatics*, vol. 7, article 326, pp. 1–17, 2006.

[172] A. Sato, H. Tichy, C. O'Huigin, P. R. Grant B, R. Grant, and J. Klein, "On the origin of Darwin's finches," *Molecular Biology and Evolution*, vol. 18, no. 3, pp. 299–311, 2001.

[173] W. Salzburger, T. Mack, E. Verheyen, and A. Meyer, "Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes," *BMC Evolutionary Biology*, vol. 5, article 17, pp. 1–15, 2005.

[174] A. L. Hughes, "Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level," *Heredity*, vol. 99, no. 4, pp. 364–373, 2007.

[175] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, New York, NY, USA, 1983.

[176] M. Lynch, *The Origins of Genome Architecture*, Sinauer Associates, Sunderland, Mass, USA, 2007.

[177] R. Nielsen, "Statistical tests of selective neutrality in the age of genomics," *Heredity*, vol. 86, no. 6, pp. 641–647, 2001.

[178] S. Aris-Brosou and L. Excoffier, "The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism," *Molecular Biology and Evolution*, vol. 13, no. 3, pp. 494–504, 1996.

[179] C. D. Bustamante, J. Wakeley, S. Sawyer, and D. L. Hartl, "Directional selection and the site-frequency spectrum," *Genetics*, vol. 159, no. 4, pp. 1779–1788, 2001.

[180] L. Zhu and C. D. Bustamante, "A composite-likelihood approach for detecting directional selection from DNA sequence data," *Genetics*, vol. 170, no. 3, pp. 1411–1421, 2005.

[181] M. Bamshad and S. P. Wooding, "Signatures of natural selection in the human genome," *Nature Reviews Genetics*, vol. 4, no. 2, pp. 99–111, 2003.

[182] M. Anisimova and D. A. Liberles, "The quest for natural selection in the age of comparative genomics," *Heredity*, vol. 99, no. 6, pp. 567–579, 2007.

[183] M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions," *Molecular Biology and Evolution*, vol. 3, no. 5, pp. 418–426, 1986.

[184] W. S. W. Wong and R. Nielsen, "Detecting selection in noncoding regions of nucleotide sequences," *Genetics*, vol. 167, no. 2, pp. 949–958, 2004.

[185] S. McCauley, S. de Groot, T. Mailund, and J. Hein, "Annotation of selection strengths in viral genomes," *Bioinformatics*, vol. 23, no. 22, pp. 2978–2986, 2007.

[186] Z. Yang, "Adaptive molecular evolution," in *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop, and C. Cannings, Eds., pp. 229–254, John Wiley & Sons, New York, NY, USA, 2nd edition, 2003.

[187] Z. Yang and R. Nielsen, "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 32–43, 2000.

[188] W. S. W. Wong, Z. Yang, N. Goldman, and R. Nielsen, "Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites," *Genetics*, vol. 168, no. 2, pp. 1041–1051, 2004.

[189] J. Zhang, R. Nielsen, and Z. Yang, "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level," *Molecular Biology and Evolution*, vol. 22, no. 12, pp. 2472–2479, 2005.

[190] J. Zhang, S. Kumar, and M. Nei, "Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes," *Molecular Biology and Evolution*, vol. 14, no. 12, pp. 1335–1338, 1997.

[191] Z. Yang, "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution," *Molecular Biology and Evolution*, vol. 15, no. 5, pp. 568–573, 1998.

[192] R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene," *Genetics*, vol. 148, no. 3, pp. 929–936, 1998.

[193] Y. Suzuki and T. Gojobori, "A method for detecting positive selection at single amino acid sites," *Molecular Biology and Evolution*, vol. 16, no. 10, pp. 1315–1328, 1999.

[194] Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites," *Genetics*, vol. 155, no. 1, pp. 431–449, 2000.

[195] T. Massingham and N. Goldman, "Detecting amino acid sites under positive selection and purifying selection," *Genetics*, vol. 169, no. 3, pp. 1753–1762, 2005.

[196] S. L. Kosakovsky Pond and S. D. W. Frost, "Not so different after all: a comparison of methods for detecting amino

acid sites under selection," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1208–1222, 2005.

[197] Z. Yang and R. Nielsen, "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages," *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 908–917, 2002.

[198] M. Anisimova and Z. Yang, "Molecular evolution of the hepatitis delta virus antigen gene: recombination or positive selection?" *Journal of Molecular Evolution*, vol. 59, no. 6, pp. 815–826, 2004.

[199] S. Aris-Brosou, "Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis," *Molecular Biology and Evolution*, vol. 22, no. 2, pp. 200–209, 2005.

[200] M. Anisimova, J. P. Bielawski, and Z. Yang, "Accuracy and power of Bayes prediction of amino acid sites under positive selection," *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 950–958, 2002.

[201] Z. Yang, W. S. W. Wong, and R. Nielsen, "Bayes empirical Bayes inference of amino acid sites under positive selection," *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 1107–1118, 2005.

[202] J. P. Huelsenbeck and K. A. Dyer, "Bayesian estimation of positively selected sites," *Journal of Molecular Evolution*, vol. 58, no. 6, pp. 661–672, 2004.

[203] S. Aris-Brosou, "Identifying sites under positive selection with uncertain parameter estimates," *Genome*, vol. 49, no. 7, pp. 767–776, 2006.

[204] M. Anisimova, R. Nielsen, and Z. Yang, "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites," *Genetics*, vol. 164, no. 3, pp. 1229–1236, 2003.

[205] M. Anisimova, J. Bielawski, K. Dunn, and Z. Yang, "Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes," *BMC Evolutionary Biology*, vol. 7, article 154, pp. 1–13, 2007.

[206] E. Zuckerkandl and L. Pauling, "Molecules as documents of evolutionary history," *Journal of Theoretical Biology*, vol. 8, no. 2, pp. 357–366, 1965.

[207] E. Zuckerkandl and L. Pauling, "Evolutionary divergence and convergence in proteins," in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds., Academic Press, New York, NY, USA, 1965.

[208] L. Bromham and D. Penny, "The modern molecular clock," *Nature Reviews Genetics*, vol. 4, no. 3, pp. 216–224, 2003.

[209] S. Aris-Brosou, "Dating phylogenies with hybrid local molecular clocks," *PLoS ONE*, vol. 2, no. 9, p. e879, 2007.

[210] H. Kishino and M. Hasegawa, "Converting distance to time: application to human evolution," *Methods in Enzymology*, vol. 183, pp. 550–570, 1990.

[211] A. Rambaut and L. Bromham, "Estimating divergence dates from molecular sequences," *Molecular Biology and Evolution*, vol. 15, no. 4, pp. 442–448, 1998.

[212] A. D. Yoder and Z. Yang, "Estimation of primate speciation dates using local molecular clocks," *Molecular Biology and Evolution*, vol. 17, no. 7, pp. 1081–1090, 2000.

[213] Z. Yang and A. D. Yoder, "Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse Lemur species," *Systematic Biology*, vol. 52, no. 5, pp. 705–716, 2003.

[214] Z. Yang, "A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times," *Acta Zoologica Sinica*, vol. 50, pp. 645–656, 2004.

[215] M. J. Sanderson, "Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach," *Molecular Biology and Evolution*, vol. 19, no. 1, pp. 101–109, 2002.

[216] A. B. Smith, D. Pisani, J. A. Mackenzie-Dodds, B. Stockley, B. L. Webster, and D. T. J. Littlewood, "Testing the molecular clock: molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata)," *Molecular Biology and Evolution*, vol. 23, no. 10, pp. 1832–1851, 2006.

[217] M. J. Sanderson, "A nonparametric approach to estimating divergence times in the absence of rate constancy," *Molecular Biology and Evolution*, vol. 14, no. 12, pp. 1218–1231, 1997.

[218] S. Aris-Brosou and Z. Yang, "Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny," *Systematic Biology*, vol. 51, no. 5, pp. 703–714, 2002.

[219] S. Aris-Brosou and Z. Yang, "Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa," *Molecular Biology and Evolution*, vol. 20, no. 12, pp. 1947–1954, 2003.

[220] S. Y. Ho, M. J. Phillips, A. J. Drummond, and A. Cooper, "Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1355–1363, 2005.

[221] J. J. Welch, E. Fontanillas, and L. Bromham, "Molecular dates for the "cambrian explosion": the influence of prior assumptions," *Systematic Biology*, vol. 54, no. 4, pp. 672–678, 2005.

[222] M. Aitkin, "Posterior Bayes factors," *Journal of the Royal Statistical Society B*, vol. 53, no. 1, pp. 111–142, 1991.

[223] A. J. Drummond, S. Y. Ho, M. J. Phillips, and A. Rambaut, "Relaxed phylogenetics and dating with confidence," *PLoS Biology*, vol. 4, no. 5, p. e88, 2006.

[224] J. P. Huelsenbeck, J. P. Bollback, and A. M. Levine, "Inferring the root of a phylogenetic tree," *Systematic Biology*, vol. 51, no. 1, pp. 32–43, 2002.

[225] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church, "Advanced sequencing technologies: methods and goals," *Nature Reviews Genetics*, vol. 5, no. 5, pp. 335–344, 2004.

[226] M. J. Moore, A. Dhingra, P. S. Soltis, et al., "Rapid and accurate pyrosequencing of angiosperm plastid genomes," *BMC Plant Biology*, vol. 6, article 17, pp. 1–13, 2006.

[227] P. Green, "2x genomes—Does depth matter?" *Genome Research*, vol. 17, no. 11, pp. 1547–1549, 2007.

[228] A. Rokas, B. L. Williams, N. King, and S. B. Carroll, "Genome-scale approaches to resolving incongruence in molecular phylogenies," *Nature*, vol. 425, no. 6960, pp. 798–804, 2003.

[229] A. G. Clark, M. B. Eisen, D. R. Smith, et al., "Evolution of genes and genomes on the *Drosophila* phylogeny," *Nature*, vol. 450, no. 7167, pp. 203–218, 2007.

[230] F. Delsuc, H. Brinkmann, and H. Philippe, "Phylogenomics and the reconstruction of the tree of life," *Nature Reviews Genetics*, vol. 6, no. 5, pp. 361–375, 2005.

[231] F. Ge, L. S. Wang, and J. Kim, "The cobweb of life revealed by genome-scale estimates of horizontal gene transfer," *PLoS Biology*, vol. 3, no. 10, p. e316, 2005.

[232] R. D. M. Page, "Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny," *Molecular Phylogenetics and Evolution*, vol. 14, no. 1, pp. 89–106, 2000.

[233] M. J. Phillips, F. Delsuc, and D. Penny, "Genome-scale phylogeny and the detection of systematic biases," *Molecular Biology and Evolution*, vol. 21, no. 7, pp. 1455–1458, 2004.

[234] H. Nishihara, N. Okada, and M. Hasegawa, "Rooting the eutherian tree: the power and pitfalls of phylogenomics," *Genome Biology*, vol. 8, no. 9, p. R199, 2007.

[235] N. Rodríguez-Ezpeleta, H. Brinkmann, B. Roure, N. Lartillot, B. F. Lang, and H. Philippe, "Detecting and overcoming systematic errors in genome-scale phylogenies," *Systematic Biology*, vol. 56, no. 3, pp. 389–399, 2007.

[236] S. B. Hedges, J. Dudley, and S. Kumar, "TimeTree: a public knowledge-base of divergence times among organisms," *Bioinformatics*, vol. 22, no. 23, pp. 2971–2972, 2006.

[237] J. E. Janečka, W. Miller, T. H. Pringle, et al., "Molecular and genomic data identify the closest living relative of primates," *Science*, vol. 318, no. 5851, pp. 792–794, 2007.

[238] S. Kumar and J. Dudley, "Bioinformatics software for biologists in the genomics era," *Bioinformatics*, vol. 23, no. 14, pp. 1713–1717, 2007.