Chapter 4

The Essentials of Computational Molecular Evolution

Stéphane Aris-Brosou and Nicolas Rodrigue

Abstract

In this chapter, we give a brief yet self-contained introduction to computational molecular evolution. In particular, we present the emergence of the use of likelihood-based methods, review the standard DNA substitution models, and introduce how model choice operates. We also present recent developments in inferring absolute dates and rates on a phylogeny and show how state-of-the-art models take inspiration from diffusion theory to link population genetics, which traditionally focuses at a taxonomic level under that of species, and molecular evolution.

Key words: Likelihood, Bayes, Model choice, Phylogenetics, Divergence times

1. Introduction

Many books (1–5) and review papers (6, 7) have been published in the recent years on the topic of computational molecular evolution, so that writing yet another primer on the very same topic may seem redundant. However, the field has undergone many changes over the last 5 years, and the models have become more sophisticated. This increase in refinement has not been motivated by a desire to complicate existing models, but rather to make an old wish come true: that of having integrated methods that can take unaligned sequences as an input, and simultaneously output the alignment, the tree, and other estimates of interest. The second driving force is more theoretical and aims at reconciling a conceptual gap between molecular evolution and population genetics.

The aim of this primer is therefore to provide readers with the essentials of computational molecular evolution, with a brief overview of recent developments. Some of the details will be left out as they are dealt with by others in this volume. Likewise, the analysis of

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*,
Methods in Molecular Biology, vol. 855, DOI 10.1007/978-1-61779-582-4_4,

genomic-scale data is briefly touched upon, but the details are left to other chapters.

2. Parsimony and Likelihood

2.1. A Brief Overview of Parsimony

The simplest phylogenetic question pertains to the reconstruction of a rooted tree with three sequences (Fig. 1). The sequences can be made of DNA, RNA, amino acids, or codons, but for the sake of simplicity we focus on DNA throughout this chapter. In the toy example below, based on ref. 8, DNA sequences are assumed to have been sampled from three different species that diverged a "long time ago." In this context, we assume that the data or gene sequences have been aligned (see Subheading 6), and that the DNA alignment is:

s_1 ATGACCCCAATACGCAAAACTAACCCCCTAATAAAATTAATTAACCACTCCTTC

s_2 ATGACCCCAATACGGAAAACTAACCCCCAAATAAAATTAATTAACCACTCATTC

s_3 ATGACGCCAATACGCAAAACTAACCGCCTAATAAAATTAATTTACCACTCATTC

The objective is to estimate which of the three fully resolved topologies in Fig. 1 is supported by the data. In order to go further, we recode the data in terms of *site patterns*, which correspond to the patterns observed in each column of our alignment. This recoding implies that columns, or sites, in our alignment evolve according to an identically and independently distributed (iid) process. With this in mind, our alignment can be recoded as follows. When all the characters (nucleotides) in a column are identical, the same letter is assigned to each character, for example x, irrespective of the actual character state. When a substitution occurs in one of the three sequences, we have three corresponding site patterns: xxy, xyx, and yxx, where the order within each site pattern respects the order of the sequences in the alignment, $s_1 s_2 s_3$.

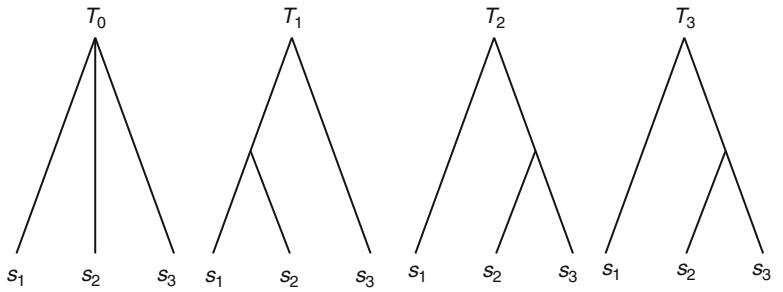

Fig. 1. The simplest phylogenetic problem. With three species, s_1, s_2, and s_3, four rooted trees are possible: T_0, the star tree, and the three resolved topologies T_1 to T_3.

Table 1
The winning-site strategy

Site pattern	Supported T_i	Count
xxx	T_0	48
xxy	T_1	3
xyx	T_2	2
yxx	T_3	1

The data alignment is reduced to a frequency table of site patterns. In the case of three sequences, only the last three site patterns are informative

s_1 xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxyxxx
s_2 xxxxxxxxxxxxxxyxxxxxxxxxxxxxxyxxxxxxxxxxxxxxxxxxxxxxxx
s_3 xxxxxyxxxxxxxxxxxxxxxxxxyxxxxxxxxxxxxxxxyxxxxxxxxxxx

The first informative site pattern, xxy, implies that sequences s_1 and s_2 at this particular site are more similar than any of these two sequences to s_3, so that this site pattern supports topology T_1, which groups sequences s_1 and s_2 together. The most intuitive idea, called the winning-site strategy, is that the topology supported by the data corresponds to the fully resolved topology that has the largest number of site patterns in its favor. In the example shown above, topology T_1 is supported by three columns (with site pattern xxy), topology T_2 by two columns (xyx), and T_3 by one column (yxx; see Table 1). This is the intuition behind parsimony, which minimizes the amount of change along a topology. Strictly speaking, unordered parsimony cannot distinguish these three trees as they all require at least one single change. Yet, it can be argued that if tree T_1 is the true tree, site pattern xxy is more likely than any other patterns as xxy requires at least one change along a long branch (the one leading to sequence s_3) while both xyx and yxx require a change along a short branch (see p. 28 *sqq.* in ref. 9; ref. 8).

A number of methodological variations exists. A very condensed overview can be found in (10), with more details in ref. 11. Most computer programs that implement substitution models where sites are iid condense the alignment as an array of site patterns; some, like PAML (12), even output these site patterns.

Note that in obtaining this topology estimate, most of the site columns were discarded from our alignment (all the xxx site patterns, representing 89% of the site in our example above). Most of our data were phylogenetically uninformative (for parsimony). We also failed to take evolutionary time into account, or any process of basic molecular biology, such as the observation that transitions (substitution of a purine [A or G] by a purine, or a pyrimidine by a

pyrimidine) are more frequent than transversions (substitution between a purine and a pyrimidine).

2.2. Assessing the Reliability of an Estimate: The Bootstrap

As with any statistical exercise estimating a quantity of interest, we would like to have a confidence interval, taken at a particular level, so that we can gauge the reliability of our estimate. A standard approach to derive confidence intervals is the bootstrap (13), a computational technique that resamples data points with replacement to simulate the distribution of any test statistic under the null hypothesis that is tested. The bootstrap, particularly useful in complicated nonparametric problems where no asymptotic results can be obtained (14), was adapted by Felsenstein to the nonstandard phylogenetic problem (15). Indeed, the problem is nonstandard in that the object for which we wish to assess accuracy is not a real-valued parameter, but a graph.

The basic idea, clearly explained in ref. 16, consists in resampling columns of the alignment, with replacement, to construct a "synthetic" alignment of the same size as the original alignment. This synthetic or bootstrap replicate is then subjected to the same tree-reconstruction algorithm used on the original data (Fig. 2). This exercise is repeated a large number of times (e.g., $\times 10^6$), and the proportion of each original bipartition (internal node) in the set of bootstrapped trees is recorded. In Fig. 2 for instance, the bipartition $s_1 s_2 | s_3$ is found in two bootstrap trees out of three, so the bootstrap support for this node is 66.7%. In this simple case with three sequences, the bootstrap support for topology T_1 is also 66.7%. This bootstrap proportion for topologies (or for *trees* when branch lengths are taken into account, in a maximum likelihood context for instance—see below) can be computed very quickly by bootstrapping not the columns of the alignment but the sitewise log-likelihood values; this bootstrap is called RELL, for "resampling estimated log-likelihood" (17).

The meaning of the bootstrap has been a matter of debate for years. As noted before (6) (see also ref. 18), the bootstrap proportion P can be seen as assessing the correctness of an internal node, and failing to do so (19), or $1 - P$ can be interpreted as a conservative probability of falsely supporting monophyly (20). Since bootstrap proportions are either too liberal or too conservative depending on the actual interpretation of P (21), it is difficult to adjust the threshold below which monophyly can be confidently ruled out (22). Alternatively, an intuitive geometric argument was proposed to explain the conservativeness of bootstrap probabilities (14) and was further developed into the Approximately Unbiased or AU test, implemented in CONSEL (23). In spite of these difficulties, the bootstrap is still widely used—and mandatory in all publications featuring a phylogeny—to assess the confidence one can have in the tree estimated from the data under a particular scheme or model (see Subheading 2.9 below).

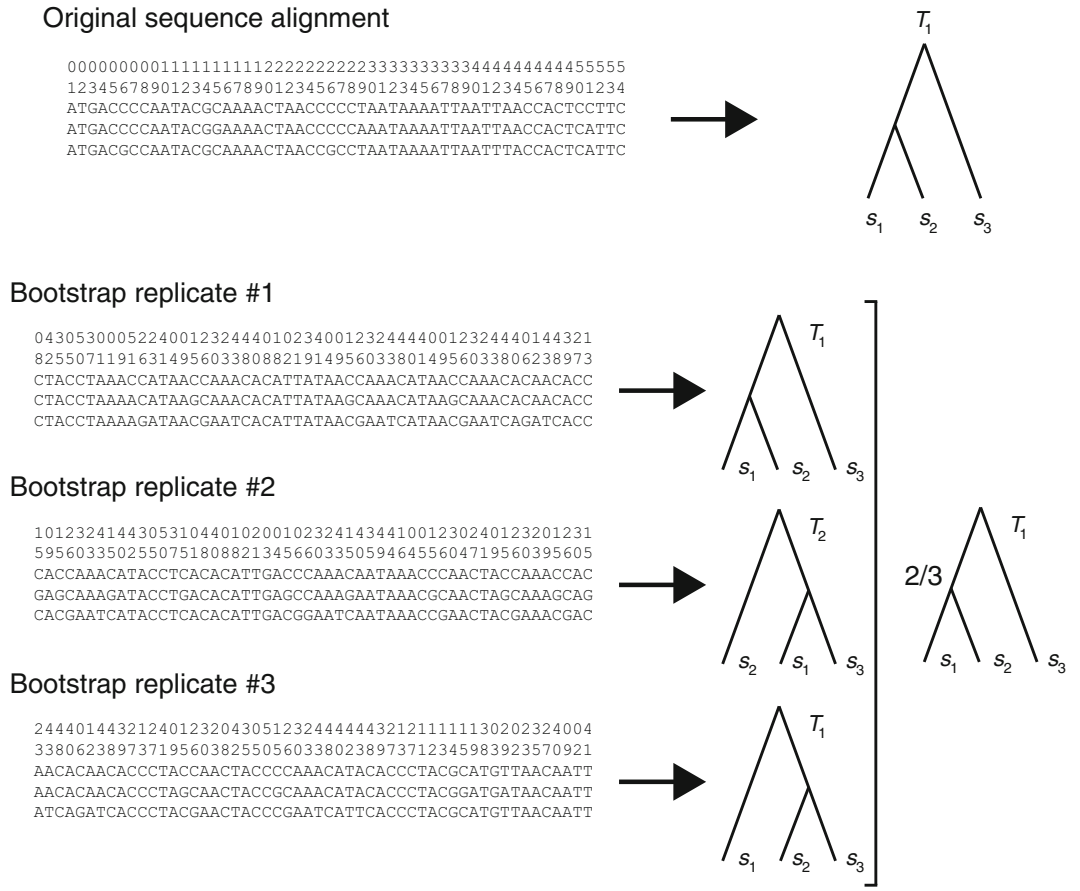

Fig. 2. The (nonparametric) bootstrap. See text for details.

2.3. Parsimony and LBA

Now that we have a means of evaluating the support for the different topologies, we can test some of the conditions under which parsimony estimates the correct tree topology. Ideally, a good method should return the correct answer with a probability of one when the number of sites increases to infinity. This desirable statistical property is called consistency. One serious criticism of parsimony is its sensitivity to long branch attraction, or LBA, even in the presence of an infinite amount of data (infinite alignment length) (24). In other words, parsimony is not statistically consistent.

Different types of model misspecification can lead to LBA, and new ones are continually identified. The topology originally used to demonstrate the artifact is represented in Fig. 3, where two long branches are separated by a shorter one. Felsenstein demonstrated that, under a simple evolutionary process, the artifact or LBA tree is reconstructed. Note that parsimony is not the only phylogenetic method affected by LBA, but because it posits a very simple model of evolution (25–27), parsimony is particularly sensitive to the artifact.

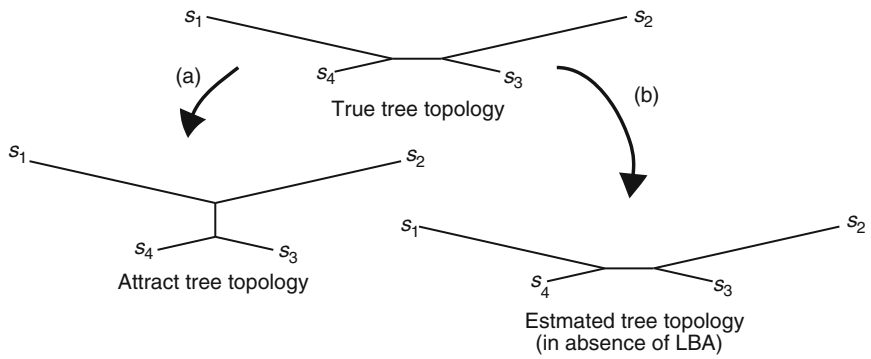

Fig. 3. The long branch attraction artifact. The true tree topology has two long branches separated by a short one. The tree reconstructed under a simple model of evolution (**a**) is the artifact or LBA tree on the *left*. The tree reconstructed under the correct model of evolution (**b**) is the correct tree, on the *right*.

The artifact has been shown to plague the analysis of numerous data sets, and a number of empirical approaches have been used to detect the artifact (28, 29). Most recent papers based on multigene analyses (e.g., refs. 30, 31) now examine carefully the effect of across-site and across-lineage rate variation (in addition to the use of heterogeneous models). For both sites and lineages, the procedure is the same and consists in successively removing either the sites that evolve the fastest, or the taxa that show the longest root-to-tip branch lengths.

2.4. Origin of the Problem

By definition, parsimony minimizes the number of changes along each branch of the tree. When there is only a small number of changes per branch, the method is expected to be accurate. However, when sequences are quite divergent, the parsimony assumption leads to underestimating the actual number of changes (Fig. 4; see also ref. 32).

Consequently, we would like a tree-reconstruction method that accounts for multiple substitutions. We would also like a method that (1) takes into account less parsimonious as well as most parsimonious state reconstructions (intervals, tests), (2) weights changes differently if they occur on branches of different length (evolutionary time), and (3) weights different kinds of events (transitions and transversions) differently (biological realism). Likelihood methods include such considerations explicitly, as they require modeling the substitution process itself.

2.5. Modeling Molecular Evolution

The basic model of DNA substitution (Fig. 5) is defined on the DNA *state space*, made of the four nucleotides thymine (T), cytosine (C), adenine (A), and guanine (G). Note that T and C are pyrimidines (biochemically, six-membered rings), while A and G are purines (fused five- and six-membered heterocyclic

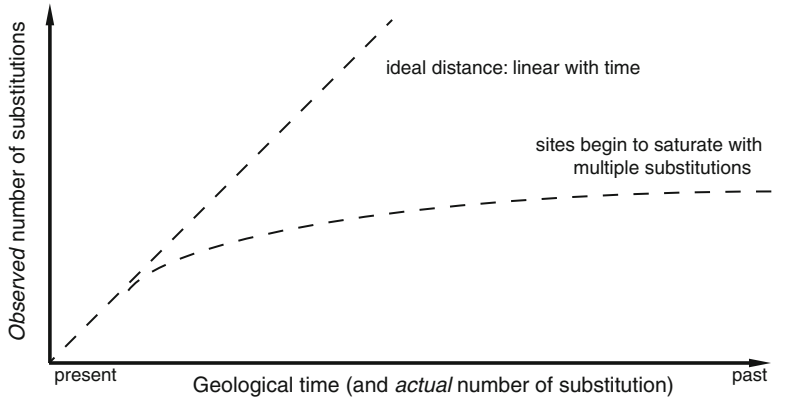

Fig. 4. Saturation of DNA sequences. As time increases, the *observed* number of differences between pairs of sequences reaches a plateau, whereas the *actual* number of substitutions keeps increasing.

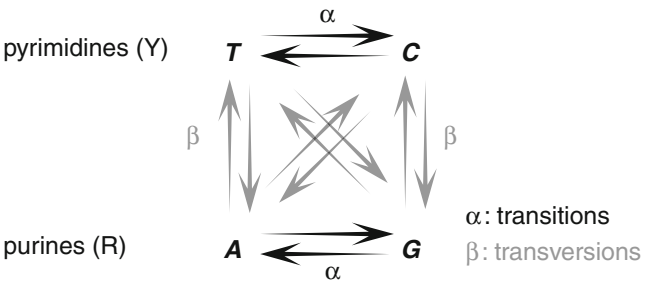

Fig. 5. Molecular evolution 101. Specification of the basic model of DNA substitution.

compounds). Depending on these two biochemical categories, two different types of substitutions can happen: transitions within a category, and transversions between categories. Their respective rates are denoted α and β in Fig. 5.

The process we want to model should describe the substitution process of the different nucleotides of a DNA sequence. Again, we will make the simplifying assumption that sites evolve under a time-homogeneous Markov process and are iid, as above. We can therefore concentrate on one single site for now (e.g., ref. 33).

At a particular site, we want to describe the change in nucleotide frequency after a short amount of time dt. For instance, the nucleotide frequency of A after dt will change from $f_A(t)$ to $f_A(t + dt)$. According to Fig. 5, $f_A(t + dt)$ will be equal to what we had at time t, $f_A(t)$, minus the quantity of A that "disappeared" by mutation during dt, plus the quantity of A that "appeared" by mutation during dt. Denoting the mutation rate as μ, the quantity of A that "disappeared" by mutation during dt is simply $f_A(t)\mu_A dt$. These

mutations away from A generated quantities of T, C, and G, in which we are not interested at the moment since we only want to know what happens to A. There are three different ways to generate A: from either T, C, or G (Fig. 5). Coming from T, mutation will generate $f_{\mathrm{T}}(t)\mu_{\mathrm{T}\to\mathrm{A}}\mathrm{d}t$ of A during dt. Similar expressions exist for C and for G, so that in total, over the three non-A nucleotides, mutation will generate $\Sigma_{i\neq\mathrm{A}}f_i(t)\mu_{i\to\mathrm{A}}\mathrm{d}t$. Mathematically, we can express these ideas as:

$$f_{\mathrm{A}}(t+\mathrm{d}t) = f_{\mathrm{A}}(t) - f_{\mathrm{A}}(t)\mu_{\mathrm{A}}\mathrm{d}t + \sum_{i\neq\mathrm{A}} f_i(t)\mu_{i\mathrm{A}}\mathrm{d}t \tag{1}$$

Equation 1 describes the change of frequency of A during a short time interval dt. Similar equations can be written for T, C, and G, so that we actually have a system of four equations describing the change in nucleotide frequencies over a short time interval dt:

$$\begin{cases} f_{\mathrm{T}}(t+\mathrm{d}t) = f_{\mathrm{T}}(t) - f_{\mathrm{T}}(t)\mu_{\mathrm{T}}\mathrm{d}t + \Sigma_{i\neq\mathrm{T}}f_i(t)\mu_{i\mathrm{T}}\mathrm{d}t \\ f_{\mathrm{C}}(t+\mathrm{d}t) = f_{\mathrm{C}}(t) - f_{\mathrm{C}}(t)\mu_{\mathrm{C}}\mathrm{d}t + \Sigma_{i\neq\mathrm{C}}f_i(t)\mu_{i\mathrm{C}}\mathrm{d}t \\ f_{\mathrm{A}}(t+\mathrm{d}t) = f_{\mathrm{A}}(t) - f_{\mathrm{A}}(t)\mu_{\mathrm{A}}\mathrm{d}t + \Sigma_{i\neq\mathrm{A}}f_i(t)\mu_{i\mathrm{A}}\mathrm{d}t \\ f_{\mathrm{G}}(t+\mathrm{d}t) = f_{\mathrm{G}}(t) - f_{\mathrm{G}}(t)\mu_{\mathrm{G}}\mathrm{d}t + \Sigma_{i\neq\mathrm{G}}f_i(t)\mu_{i\mathrm{G}}\mathrm{d}t \end{cases} \tag{2}$$

which, in matrix notation, can simply be rewritten as:

$$F(t+\mathrm{d}t) = F(t) + QF(t)\mathrm{d}t \tag{3}$$

with an obvious notation for F, while the *instantaneous rate matrix* Q is:

$$Q = \begin{pmatrix} -\mu_{\mathrm{T}} & \mu_{\mathrm{TC}} & \mu_{\mathrm{TA}} & \mu_{\mathrm{TG}} \\ \mu_{\mathrm{CT}} & -\mu_{\mathrm{C}} & \mu_{\mathrm{CA}} & \mu_{\mathrm{CG}} \\ \mu_{\mathrm{AT}} & \mu_{\mathrm{AC}} & -\mu_{\mathrm{A}} & \mu_{\mathrm{AG}} \\ \mu_{\mathrm{GT}} & \mu_{\mathrm{GC}} & \mu_{\mathrm{GA}} & -\mu_{\mathrm{G}} \end{pmatrix} \tag{4}$$

In all the following matrices, we will use the same order for nucleotide: T, C, A, and G, which follows the order in which codon tables are usually written. Recall that μ_{ij} is the mutation rate from nucleotide i to nucleotide j. Note also that the sum of each row is 0.

Let us rearrange the matrix notation from Eq. 3 as:

$$F(t+\mathrm{d}t) - F(t) = QF(t)\mathrm{d}t \tag{5}$$

and take the variation limit when $\mathrm{d}t \to 0$:

$$\frac{\mathrm{d}F(t)}{\mathrm{d}t} = QF(t) \tag{6}$$

which is a first order differential equation that can be integrated as:

$$F(t) = e^{Qt}F(0) \tag{7}$$

Very often, this last Eq. 7 is written as $F(t) = P(t)F(0)$, where $F(0)$ is conveniently taken to be the identity matrix and $P(t) = \{P_{i,j}(t)\} = e^{Qt}$ is the matrix of probabilities of going from state i to j during a finite time duration t. Note that the right-hand side of this equation is a matrix exponentiation, which is not the same as the exponential of all the elements (row and columns) of that matrix. The computation of the term e^{Qt} demands that a spectral decomposition of the matrix Q be realized. This means finding a diagonal matrix D of eigenvalues and a matrix M of (right) eigenvectors so that:

$$P(t) = Me^{Dt}M^{-1} \tag{8}$$

The exponential of the diagonal matrix D is simply the exponential of the diagonal terms.

Except in the simplest models of evolution, finding analytical solutions for the eigenvalues and associated eigenvectors can be tedious. As a result, numerical procedures are employed to solve Eq. 8. Alternatively, a Taylor expansion can be used to *approximate* $P(t)$.

If all entries in Q are positive, any state or nucleotide can be reached from any other in a finite number of steps (all states "communicate") and the base frequencies have a stationary distribution $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$. This is the steady state reached after an "infinite" amount of time, or long enough for the Markov process to forget its initial state, starting from "random" base frequencies.

2.6. Computation on a Tree

Now that we know how to determine the rate of change of nucleotide frequencies during a time interval dt, we can compute the probability of a particular nucleotide change on a tree. The simplest case, though somewhat artificial with only two sequences, is depicted in Fig. 6.

We are looking at a particular nucleotide position, denoted j, for two aligned sequences. The observed nucleotides at this position are T in sequence 1, and C in sequence 2. The branch separating T from C has a total length of $t_0 + t_1$. For the sake of convenience, we set an

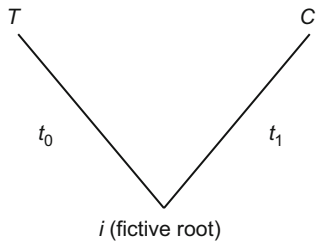

Fig. 6. Likelihood computation on a small tree. See text for details.

arbitrary root along this path. The likelihood at site j is then given by the probability of going from the fictive root i to T in t_0, and from i to C in t_1. Any of the four nucleotides can be present at the fictive root. As we do not know which one was there, we sum these probabilities over all possible state, weighted by their prior probabilities, the equilibrium frequencies π_i. In all, we have the likelihood ℓ_j at site j:

$$\ell_j = \sum_{i=\{\mathrm{T,C,A,G}\}} \pi_i P_{i,T}(t_0) P_{i,C}(t_1) \tag{9}$$

which is equivalent to the Chapman–Kolmogorov equation (34). As all the sites are assumed to be iid, the likelihood of an alignment is the product of the site likelihoods in Eq. 9.

Note that this example is somewhat artificial: with only two sequences, we can compute the likelihood directly with $\pi_{\mathrm{T}} P_{\mathrm{T,C}}(t_0 + t_1) = \pi_{\mathrm{C}} P_{\mathrm{C,T}}(t_0 + t_1)$; the full summation over unknown states as in Eq. 9 is required with three sequences or more. When analyzing a multiple-sequence alignment of S sequences, there will be many nodes in the tree for which the character state is unknown, which means that the summation required will involve many terms. Specifically, the sum will be over 4^{S-3} terms. Fortunately, terms can be factored out of the summation, and a dynamic programming algorithm in $4^2 S$, called the pruning algorithm (35), can be used (see ref. 11 for details).

2.7. Substitution Models and Instantaneous Rate Matrices Q

Now that we have almost all the elements to compute the likelihood of a set of parameters, including the tree (branch lengths + topology; see Subheading 2.10), the only missing element required to compute the likelihood at each site, as in Eq. 9 for instance, is the specification of the instantaneous rate matrix Q as in Eq. 4. Remember that the $\mu_{i,j}$ represent mutation rates from state (nucleotide) i to j. This matrix is generally rewritten as:

$$Q = \mu \begin{pmatrix} - & r_{\mathrm{TC}} & r_{\mathrm{TA}} & r_{\mathrm{TG}} \\ r_{\mathrm{CT}} & - & r_{\mathrm{CA}} & r_{\mathrm{CG}} \\ r_{\mathrm{AT}} & r_{\mathrm{AC}} & - & r_{\mathrm{AG}} \\ r_{\mathrm{GT}} & r_{\mathrm{GC}} & r_{\mathrm{GA}} & - \end{pmatrix} \tag{10}$$

so that each entry r_{ij} is a rate of change from nucleotide i to nucleotide j. The diagonal entries are left out, indicated by a "–," and are in fact calculated as the negative sum of the off-diagonal entries (as rows sum to 0).

The simplest specification of Q would be that all rates of change are identical, so that Q becomes (leaving out the mutation rate μ and indexing the matrix to indicate the difference):

$$Q_{\mathrm{JC}} = \begin{pmatrix} - & 1 & 1 & 1 \\ 1 & - & 1 & 1 \\ 1 & 1 & - & 1 \\ 1 & 1 & 1 & - \end{pmatrix} \tag{11}$$

which is the model proposed by Jukes and Cantor (36) and often noted "JC" or "JC69." Under the specification of Eq. 11, this model has no free parameter. The process is generally scaled such that the unit of branch lengths can be interpreted as an expected number of substitutions per site.

Of course, this model is extremely simplistic and neglects a fair amount of basic molecular biology. In particular, it overlooks two observations. First, base frequencies are not all equal in actual DNA sequences, but are rather skewed, and second, transitions are more frequent than transversions (see Subheading 2.5).

The way to account for this first "biological realism" is as follows. If DNA sequences were made exclusively of "A"s for instance, that would mean that all mutations are towards the observed base, in this case A, whose equilibrium or stationary frequency is π_{A}. The same reasoning can be used for arbitrary equilibrium frequencies π, so that all relative rates of change in Q become proportional to the vector of equilibrium frequency π of the *target* nucleotide. In other words, the instantaneous rate matrix Q becomes:

$$Q_{\mathrm{F81}} = \begin{pmatrix} - & \pi_{\mathrm{C}} & \pi_{\mathrm{A}} & \pi_{\mathrm{G}} \\ \pi_{\mathrm{T}} & - & \pi_{\mathrm{A}} & \pi_{\mathrm{G}} \\ \pi_{\mathrm{T}} & \pi_{\mathrm{C}} & - & \pi_{\mathrm{G}} \\ \pi_{\mathrm{T}} & \pi_{\mathrm{C}} & \pi_{\mathrm{A}} & - \end{pmatrix} \tag{12}$$

again with the requirement that rows sum to 0. This matrix represents the Felsenstein or F81 model (35). This model has four parameters (the four base frequencies), but since base frequencies sum to 1, we only have three *free* parameters.

The second "biological realism," accounting for the different rates of transversions and transitions, can be described by saying that transitions occur κ times faster than transversions. From Fig. 5, recall that transitions are mutations from T to C (and vice versa) and from A to G (and vice versa). This translates into:

$$Q_{\mathrm{K80}} = \begin{pmatrix} - & \kappa & 1 & 1 \\ \kappa & - & 1 & 1 \\ 1 & 1 & - & \kappa \\ 1 & 1 & \kappa & - \end{pmatrix} \tag{13}$$

This model is called the Kimura two-parameter model or K80 (or K2P) (37). The model is alternatively described with the two rates α and β (see Fig. 5). In the "κ version" of the model as in Eq. 13, there is only one free parameter.

Of course it is possible to account for both kinds of "biological realisms," unequal equilibrium base frequencies and transition bias, all in the same model, whose generator Q becomes:

$$Q_{\mathrm{HKY}} = \begin{pmatrix} - & \pi_{\mathrm{C}}\kappa & \pi_{\mathrm{A}} & \pi_{\mathrm{G}} \\ \pi_{\mathrm{T}}\kappa & - & \pi_{\mathrm{A}} & \pi_{\mathrm{G}} \\ \pi_{\mathrm{T}} & \pi_{\mathrm{C}} & - & \pi_{\mathrm{G}}\kappa \\ \pi_{\mathrm{T}} & \pi_{\mathrm{C}} & \pi_{\mathrm{A}}\kappa & - \end{pmatrix} \tag{14}$$

which corresponds to the Hasegawa Kishino Yano or HKY (or HKY85) model (38). This model has four free parameters: κ and three base frequencies.

The level of "sophistication" goes "up to" the General Time-Reversible model (39), denoted GTR or REV, which has for generator:

$$Q_{\mathrm{GTR}} = \begin{pmatrix} - & a\pi_{\mathrm{C}} & b\pi_{\mathrm{A}} & c\pi_{\mathrm{G}} \\ a\pi_{\mathrm{T}} & - & d\pi_{\mathrm{A}} & e\pi_{\mathrm{G}} \\ b\pi_{\mathrm{T}} & d\pi_{\mathrm{C}} & - & \pi_{\mathrm{C}} \\ c\pi_{\mathrm{T}} & e\pi_{\mathrm{C}} & \pi_{\mathrm{A}} & - \end{pmatrix} \tag{15}$$

The number of free parameters is now eight (three base frequencies plus five nucleotide propensities). The name is derived from the time-reversibility constraint, which implies that the likelihood is independent of the actual orientation of time.

In fact, there exists only a few "named" additional substitution models (11), most of which are time-reversible models, while a total of 203 models can be derived from GTR (40). We have focused solely on DNA models in this chapter, but the problem is similar with amino acid or codon models, except that the number of parameters increases quickly. We have also limited ourselves to time-reversible time-homogeneous models, but irreversible non-homogeneous models were developed some time ago (41) and are used, for instance, to root phylogenies (42) or to help alleviate the effects of LBA (31).

2.8. Some Computational Aspects

2.8.1. Optimization of the Likelihood Function

For a given substitution model, how should parameters be estimated, given the (potentially) high dimensionality of the model? Analytical solutions consist in determining when the first derivative of the likelihood function is equal to zero (with a change of sign in the second derivative). However, finding the root of the likelihood function analytically is only possible in the simple case of three sequences of binary characters under the assumption of the molecular clock (see Subheading 3.1) (8). As a result, numerical solutions must be found to maximize the likelihood function.

A number of ideas have been combined to search efficiently for the parameter values that maximize the likelihood function. Most programs will start from a random starting point, for example $\left(\theta_1^{(0)}, \theta_2^{(0)}\right)$, denoted by an x in Fig. 7, where we limit ourselves

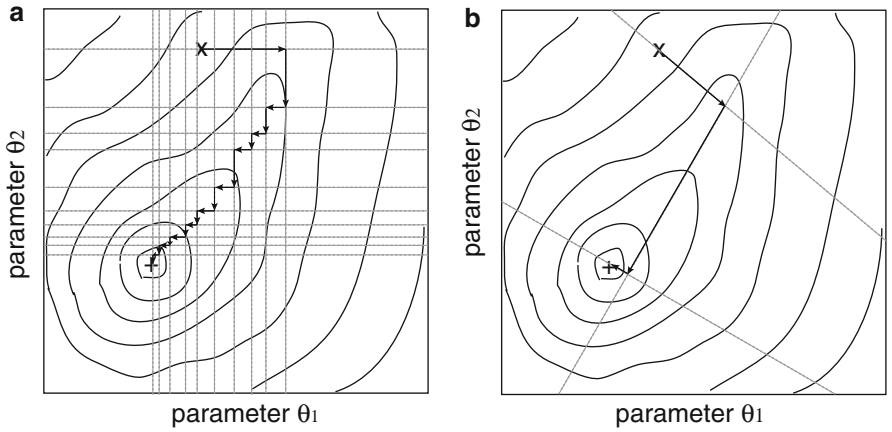

Fig. 7. Two optimization strategies. The likelihood surface of a function with two parameters θ_1 and θ_2 (e.g., two branch lengths) is depicted as a contour plot, whose highest peak is at the + sign. (**a**) Optimization of one parameter at a time. (**b**) Optimization of all parameters simultaneously. See text for details.

to a two-parameter example. The optimization procedure can follow one of two strategies. In the first one, parameters are optimized one at a time. In Fig. 7a, parameter θ_1 is first optimized to maximize the likelihood function with a line search, which defines a direction along which the other parameter (θ_2) or parameters in the multidimensional case are kept constant. Once $\theta_1^{(1)}$ is found, a new direction is defined to optimize θ_2, and so on so forth until convergence to the maximum of the likelihood function. As shown in Fig. 7a, many iterations can be required, in particular when the parameters θ_1 and θ_2 are correlated. The alternative to optimizing one parameter at a time is to optimize all parameters simultaneously. In this case (Fig. 7b), an initial direction is defined at $\left(\theta_1^{(0)}, \theta_2^{(0)}\right)$ such that the slope at this point is maximized. The process is repeated until convergence. More technical details can be found in ref. 3. The simultaneous optimization procedure generally requires fewer steps than optimizing parameters one at a time, but not always. Since the computation of the likelihood function is the most expensive computation of these algorithms, the simultaneous optimization is much more efficient, at least in our toy example.

How general is this result? Simultaneously, optimizing parameters of the substitution model, while optimizing branch lengths one at a time, was shown to be more effective on large data sets (43), potentially because of the correlation that exists between some of the parameters entering the Q matrix (see Subheading 2.7).

2.8.2. Convergence

Convergence is usually reached either when the increment in the log-likelihood score becomes smaller than an ϵ value, usually set to a small number such as 10^{-6} (but yet a number larger than the

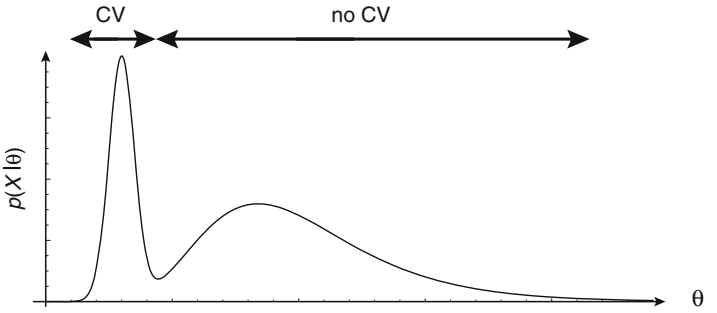

Fig. 8. Likelihood surfaces behaving badly. Schematic of the probability surface of the function $p(X|\theta)$ is plotted as a function of θ. Most line search strategies will converge (CV) to the MLE when the initial value is in the "CV" interval, and fail when it is in the "no CV" interval. Adapted with permission from ref. 168.

machine ϵ), or when the log-likelihood score has not changed after a predetermined number of iterations. None of these stopping rules, however, guarantees that the global maximum of the likelihood function has been found. Therefore, it is generally recommended to run the optimization procedure at least twice, starting from different points in the parameter space, and to check that the likelihood score after optimization is the same across the different runs (Fig. 8). If this is not the case, additional runs may be required, and the one with the largest likelihood is chosen for inference (e.g., ref. 44).

In many instances though, different substitution models will give different tree topologies, and therefore different biological conclusions. One difficulty is therefore to know which model should be used to analyze a particular data set.

2.9. Selection of the Appropriate Substitution Model

One important issue in model selection is about the trade-off between bias and variance (45): a simple model will fail to capture all the sophistication of the actual substitution process, and will therefore be highly biased even if all the parameters can be estimated with tight precision (little variance). Alternatively, a highly parameterized model will "spread" the information available from the data over a large number of parameters, hereby making their estimation difficult (flat likelihood surface; see Subheading 2.8), with a large variance, in spite of perhaps being a more realistic model with less bias. The objective of most model selection procedure is therefore to find not the *best* model in terms of likelihood score, but the *most appropriate* model, the one that strikes the right balance between bias and variance in terms of number of parameters. However, we argue that optimizing for this bias-variance trade-off works only for statistical procedures, be they for instance frequentist (LRT: likelihood ratio test) or Bayesian (BF: Bayes factor), while information-theoretic criteria (e.g., AIC: Akaike

information criterion) aim at selecting the model that is *approximately* closest to the "true" biological process.

The bias-variance trade-off mainly concerns the comparison of models that are based on the same underlying rationale, for instance choosing among the 203 models that can be derived from GTR. We may also be interested in comparing models that are based on very different rationales. The LRT is suited for assessing the bias-variance trade-off, while Bayesian approaches and cross-validation (CV) can be used for more general model comparisons. Here, we review four approaches to model selection: LRT, BF, AIC, and CV.

2.9.1. The Likelihood Ratio Test

The substitution models presented above have one key property: it is possible to reduce the most sophisticated time-reversible named model (GTR + Γ + I) to any simpler model by imposing some constraints on parameters. As a result, the models are said to be nested, and statistical theory (the Neyman–Pearson lemma) tells us that there is an optimal (most powerful) way of comparing two nested models (a simple null vs. a simple alternative hypothesis) based on the LRT.

The test statistic of the LRT is twice the log-likelihood difference between the most sophisticated model (which by definition is always the one with the highest likelihood—if this is not the case, there is a convergence issue; see Subheading 2.8) and the simpler model. This test statistic follows asymptotically a χ^2 distribution (under certain regularity conditions), and the degree of freedom of the test is equal to the difference in the number of free parameters between the two models.

The null hypothesis is that the two competing models explain the data equally well. The alternative is that the most sophisticated model explains the data better than the simpler model. If the null hypothesis cannot be rejected at a certain level (type-I error rate), then, based on the argument developed above, the simpler model should be used to analyze the data. Otherwise, if the null hypothesis can be rejected, the more sophisticated model should be used to analyze the data. Note that a test never leads to accepting a null hypothesis; the only outcomes of a test are either *reject*, or *fail to reject* a null hypothesis.

Intuitively, we can see the null hypothesis H_0 as stating that a certain parameter θ is equal to θ_0. The maximum likelihood estimate (MLE) is at $\hat{\theta}$, which is our alternative hypothesis H_1, left unspecified. We note the log-likelihood as $\ln\, p(X|\theta) = \ell(\theta)$, where X represents the data. Under H_0, we have $\theta = \theta_0$, while under H_1 we have $\theta = \hat{\theta}$. The log-likelihood ratio is therefore $\ln \mathrm{LR} = \ell(\hat{\theta}) - \ell(\theta_0)$. Under the null H_0, $\ell(\hat{\theta}) = 0$ (by definition). The log-likelihood ratio then reduces to $\ln\ \mathrm{LR} = -\ell(\theta_0)$. We can then take the Taylor expansion of the log-likelihood function ℓ around $\hat{\theta}$, which gives us $\ell \approx \frac{1}{2}(\hat{\theta} - \theta_0)^2 \dfrac{d^2\ell}{\mathrm{d}\theta^2}$ (recall

that $\ell(\hat{\theta}) = 0$, so that the first terms of the series "disappear"). Therefore, log-likelihood ratio can be approximated by $-\frac{1}{2}(\hat{\theta} - \theta_0)^2 \frac{d^2\ell}{d\theta^2}$. Recall that Fisher's information is negative the reciprocal of the second derivative of the likelihood function, so that:

$$\ln \mathrm{LR} \approx \frac{(1/2)\left(\hat{\theta} - \theta_0\right)^2}{\mathrm{var}(\theta_0)} \tag{16}$$

which follows asymptotically half a χ^2 distribution. Hence the usual approximation:

$$2 \ln \mathrm{LR} = 2 \times (\ell_1 - \ell_0) \sim \chi^2_k \tag{17}$$

with k being the difference in the number of free parameters between the two models 0 and 1. The important points in this intuitive outline of the proof are that (1) the two hypotheses need to be nested and (2) taking the Taylor expansion around $\hat{\theta}$ requires that the likelihood function be continuous at that point, which implies that ℓ is differentiable left and right of $\hat{\theta}$. Therefore, testing points at the boundary of the parameter space cannot be done by approximating the distribution of the test statistic of the LRT by a regular χ^2 distribution, as noted many times in molecular evolution (46–54). A solution still involves the LRT, but the asymptotic distribution becomes a mixture of χ^2 distributions (55).

An approach that has become popular under the widespread adoption of computer programs such as ModelTest (56) and jModelTest (57) is the hierarchical LRT or hLRT. This hierarchy goes from the simplest model (JC) to the set of most complex models ($+\Gamma$ + I), traversing a tree of models. The issue is that there is more than one way to traverse this tree of models, and that depending on which way is adopted, the procedure may end up selecting different models (58, 59).

2.9.2. Information Theoretic Approaches

Information theory provides us with a number of solutions to circumvent the three limitations of the LRT (nestedness, continuity, and dependency on the order in which models are compared).

The core of the information-based approach is the Kullback–Leibler (KL) distance, or information (60), which measures the distance between an approximating model g and a "true" model f (45). This distance is computed as:

$$d_{\mathrm{KL}}(f, g) = \int f(x) \ln \frac{f(x)}{g(x|\theta)} \mathrm{d}x \tag{18}$$

where θ is a vector of parameters entering the approximating model g, and x represents the data. Note that this distance is not symmetric, as typically $d_{\mathrm{KL}}(f, g) \neq d_{\mathrm{KL}}(g, f)$, and that the "true" model f is unknown. The idea is to rewrite $d_{\mathrm{KL}}(f, g)$ in a slightly different

form, to make it clear that Eq. 18 is actually a difference between two expectations, both taken with respect to the unknown "truth" f:

$$d_{\mathrm{KL}}(f, g) = E_f[f(x)\ \ln f(x)] - E_f[f(x)\ \ln g(x|\theta)] \qquad (19)$$

Equation 19 therefore measures the loss of information incurred by fitting g when the data x actually come from f. As f is unknown, $d_{\mathrm{KL}}(f, g)$ cannot be computed as such.

Two points are key to deriving the criterion proposed by Akaike (see ref. 45). First, we usually want to compare at least two approximating models, g_0 and g_1. We can then measure which one is closest to the "true" process f by taking the difference between their respective KL distances. In the process, the direct reference to the "true" process cancels out. As a result, the "best" model among g_0 and g_1 is the one that is closest to the "true" process f: it is the model that *minimizes* the distance to f. By setting model parameters to their MLEs, we now deal with *estimated* distances, but these are still with respect to the unknown f.

Second, in the context of a frequentist approach, we would repeat the experiment of sampling data an infinite number of times. We would then compute the *expected estimated KL distance*, so that model selection can be done on the sole estimated log-likelihood value. Akaike, however, showed that this latter approximation is biased, and must be adjusted by a term that is *approximately equal* to the number of parameters K entering model g (see ref. 45). For "historical reasons" (similarity with asymptotic theory with the normal distribution), the selection criterion is multiplied by 2 to give the well-known definition of the AIC:

$$AIC = -2 \ln \ell(\hat{\theta}) + 2K \qquad (20)$$

Unlike the case of the hLRT, where we were selecting the "most appropriate model" (with respect to the bias-variance trade-off), in the case of AIC we can select the *best* model. This best model is the one that is closest to the "true" unknown model (f), with the smallest relative estimated expected KL distance. The best AIC model therefore minimizes the criterion in Eq. 20.

A small-sample second-order version of AIC exists, where the penalty for extra parameters ($2K$ in Eq. 20) is slightly modified to account for the trade-off between information content in the data and K (see ref. 45). In our experience, we find it advisable to use this small-sample correction irrespective of the actual size of the data, since this correction vanishes in large and informative samples, but corrects for proper model ranking when K becomes very large compared to the amount of information (e.g., in phylogenomics where models are partitioned with respect to hundreds of genes).

The AIC has been shown to tend to favor parameter-rich models (61–65), which has motivated the use and development of alternative approaches in computational molecular evolution. These include, the Bayesian Information Criterion (66), and the

decision theory or DT approach, which is based on ΔAIC weighted by squared branch length differences (61). Most of these approaches, including the hLRT, have recently been compared in a simulation study that suggests, in agreement with empirical studies (62, 67), that both BIC and DT have the highest accuracy and precision (65).

Note finally that all these approaches are not limited to selecting the most appropriate or the best model of evolution. Disregarding the hLRT, which requires that models be nested (to be able to use the χ^2 approximation; otherwise, see ref. 55), AIC, BIC, etc. allow us to compare nonnested models and, in particular, phylogenetic trees (branch lengths plus topology).

2.9.3. The Bayesian Approach

The Bayesian framework has permitted the development of two main approaches, which are actually two sides of the same coin: one based on finding the model that is the most probable a posteriori, and one based on ranking models and estimating a quantity called the BF.

In a nutshell, the frequentist approaches developed in the previous sections are based on the likelihood, which is the probability of the data, given the parameters: $p(X|\theta)$. However, this approach may not be the most intuitive, since most practitioners are not interested in knowing the conditional probability of their data, as the data were collected to learn more about the processes that generated them. It can therefore be argued that the Bayesian approach, which considers the probability of the parameters given the data or $p(\theta|X)$, is more intuitive than the frequentist approach. Unlike likelihood, which relies on the function $p(X|\theta)$ and permits point estimation, Bayesian inference is based on the posterior distribution $p(\theta|X)$. This distribution is often summarized by a centrality measure such as its mode, mean, or median. Measures of uncertainty are based on *credibility* intervals, the Bayesian equivalent of *confidence* intervals. Typically, credibility intervals are taken at the 95% cutoff and are called highest posterior densities (HPDs).

The connection between posterior probability and likelihood is made with Bayes' inversion formula, also called Bayes' theorem, by means of a quantity called the prior distribution $p(\theta)$:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \tag{21}$$

The prior represents what we think about the process that generated the data, before analyzing the data, and is at the origin of all controversies surrounding Bayesian inference. In practice, priors are more typically chosen based on statistical convenience, and often have nothing to do with our genuine state of knowledge about parameters before observing the available data. We will see in Subheading 3.1 that priors can be used to distinguish between

parameters that are confounded in a maximum likelihood analysis (model), so that we argue that the frequentist versus Bayesian controversy is sterile, and we advocate a more pragmatic approach, that often results in the mixing of both approaches (in their concepts and techniques) (68, 69).

All models have parameters. Subheading 2.7 treats substitution models, which can have nine free parameters in the case of GTR + Γ. Most people are not really interested in these parameters θ or in their estimates $\hat{\theta}$, but have to use them in order to estimate a phylogenetic tree τ. These parameters θ are called *nuisance* parameters because they enter the model but are not the focus of inference. The likelihood solution consists in setting these parameters to their MLE, ignoring the uncertainty with which they can be estimated, while the Bayesian approach will integrate them out, directly accounting for their uncertainty:

$$p(X|\tau) = \int_{\Theta} p(X|\tau,\theta)\ p(\theta)\ \mathrm{d}\theta \tag{22}$$

One difficulty in Bayesian inference is about the denominator in Eq. 21, as this denominator often has no analytical solution. In spite of being a normalizing constant, $p(X)$ requires integrating out nuisance parameters by means of prior distributions as in Eq. 22. Thus, it is easy to see from Eq. 21 that the posterior distribution of the variable of interest (e.g., τ) can quickly become complicated:

$$p(\tau|X) = \int_{\Theta} \frac{p(X|\tau,\theta)p(\tau)p(\theta)}{\Sigma_T p(X|\tau,\theta)p(\tau)p(\theta)}\mathrm{d}\theta \tag{23}$$

where τ and θ are assumed to be independent and the discrete sum is taken over the set T of all possible topologies (see Subheading 2.10). However, the ratio of posteriors evaluated at two different points will simplify: as the denominator in Eq. 23 is a constant, it will cancel out from the ratio. This simple observation is at the origin of an integration technique for approximating the posterior distribution in Eq. 23: Markov chain Monte Carlo (MCMC) samplers. A very clear introduction can be found in ref. 70.

Building on this, two approaches can be formulated to compare models in a Bayesian framework. The first is to treat the model as a "random variable," and compute its posterior probability. The *best* model is then the one that has the highest posterior probability. This approach is typically implemented in a reversible-jump MCMC (or rjMCMC) sampler (e.g., see ref. 40).

The alternative is to use the Bayesian equivalent of the LRT, the BF. Rather than comparing two likelihoods, the BF compares the probability of the data under two models, M_0 and M_1:

$$\mathrm{BF}_{0,1} = \frac{p(X|M_0)}{p(X|M_1)} \tag{24}$$

More specifically, $BF_{0,1}$ evaluates the weight of evidence in favor of model M_0 against model M_1, with $BF_{0,1} > 1$ considered as evidence in favor of M_0. Just as in a frequentist context, where a null hypothesis is significantly rejected at a certain threshold, 5%, 1%, or less depending on different costs or error types, BFs can be evaluated on a specific scale (71). However, because this scale is just as ad hoc as in a frequentist setting, it might be preferable to use the probability of the data under a particular model $p(X|M_i)$ as a means of ranking models M_i.

The quantity $p(X|M_0)$, which is the denominator in Eq. 23 (where we did not include the dependence on the model in the notation), is called the marginal likelihood. Note that it is also an expectation with respect to a prior probability distribution:

$$p(X|M_0) = \int_{\Theta} p(X|\theta, M_0)p(\theta|M_0)\ \mathrm{d}\theta \tag{25}$$

A number of approximations to evaluate Eq. 25 exist and are reviewed in ref. 72 (see also refs. 73, 74). The simplest one is based on the harmonic mean of the likelihood sampled from the posterior distribution (75). The way this estimator is derived demands to understand how integrals can be approximated. Briefly, to compute $I = \int g(\theta)p(\theta)\ \mathrm{d}\theta$, generate a sample from a distribution $p^*(\theta)$ and calculate the simulation-consistent estimator $I = \Sigma\ w_i g(\theta)/\Sigma w_i$, where w_i is the *importance function* $p(\theta)/p^*(\theta)$. Take $g = p(X|\theta)$ and $p^*(\theta) = p(X|\theta)p(\theta)/p(X)$, then $\hat{I} = \hat{p}(X|M_0) = \lim_{N\to\infty}\left(\frac{1}{N}\sum\frac{1}{p(X|\theta_i)}\right)^{-1}$ with $\theta \sim p(\theta|X)$ (see Supplementary information in ref. 76). As a result, a very simple way to estimate the marginal likelihood and BFs is to take the output of an MCMC sampler and compute the harmonic mean of the likelihood values (not the log-likelihood values) sampled from the posterior distribution.

Because of its simplicity, this estimator is now implemented in most popular programs such as MrBayes (77) or BEAST (78). However, it might be considered as the worst estimator possible, because its results are unstable (75, 79) and biased towards the selection of parameter-rich models (73). An alternative and reliable estimator, based on thermodynamic integration (TI; ref. 73), is much more demanding in terms of computation. Indeed, it requires running MCMC samplers morphing one model into the other (and vice versa), which can increase computation time by up to an order of magnitude (73). Improvements of the TI estimator are however available. The Stepping-Stone approach builds on importance sampling and TI to speed up the computation while maintaining the accuracy of the standard TI estimator (74, 80).

2.9.4. Cross-Validation

Cross-validation is another model selection approach, which is extremely versatile in that it can be used to compare any set of models of interest. Besides, the approach is very intuitive. In its simplest form, cross-validation consists in dividing the available data into two sets, one used for "training" and the other one used for "validating." In the training step (TS), the model of interest is fitted to the training data in order to obtain a set of MLEs. These MLEs are then used to compute the likelihood using the validation data (validation step: VS). Because the validation data were not part of the training data, the likelihood values computed during VS can be directly used to compare models, without requiring any explicit correction for model dimensionality.

The robustness of the cross-validation scores can be explored in various ways, such as repeating the above procedure with a switched labeling of training and validation data (hence the expression *cross-validation*). Of course, this simple twofold cross-validation could be extended to n-fold cross-validation, where the data are subdivided into n subsets, with $n - 1$ subsets serving for training, and one for validation. Ideally, the procedure is repeated $n - 1$ additional times.

We know of only two examples of its use in phylogenetics, one in the ML framework (81) and one with a Bayesian approach (82). Given the increasing size of modern data sets, putting aside some of the data for validation is probably not going to dramatically affect the information content of the whole data set. As a result, model selection via cross-validation, which is statistically sound, could become a very popular approach.

2.10. Finding the Best Tree Topology

2.10.1. Counting Trees

Now that we can select a model of evolution (Subheading 2.9) and estimate model parameters (Subheading 2.8) under a particular model (Subheading 2.5), how do we find the optimal tree? The toy example in Subheading 2.1 suggested that we score all possible tree topologies and choose for inference the one that has the highest score. However, a simple counting exercise shows that an exhaustive examination of all possible topologies is not realistic.

Figure 9 shows how to count tree topologies. Starting from the simplest possible unrooted tree, with three taxa, there are three positions where a fourth branch (leading to a fourth taxon) can be added. As a result, there are three possible topologies with four taxa. For each of these, there are four places on the tree where a fifth branch can be added, which leads to a total of $3 \times 5 = 15$ topologies with five taxa. A recursion appears immediately, and it can be shown that the total number of unrooted topologies with n taxa is equal to $1 \times 3 \times \cdots \times 2n - 5$ (83) (see ref. 11 for the deeper history), which, as given in ref. 84, is equal to:

$$N_{unrooted}^{T(n)} = \frac{(2n-5)!}{2^{n-3}(n-3)!} = \frac{2^{n-2}\Gamma(n-(3/2))}{\sqrt{\pi}} \qquad (26)$$

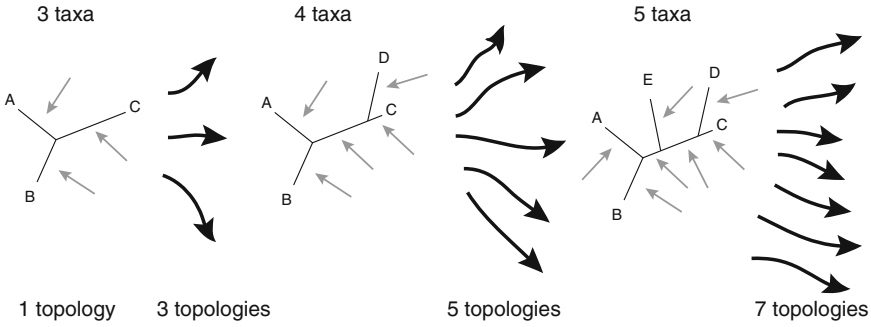

Fig. 9. Procedure to count the number of unrooted topologies. The *top line* shows the current number of taxa included in the tree below. *Gray arrows* indicate locations where an additional branch can be grafted to add one taxon. *Black arrows* show the resulting number of topologies after addition of a branch (taxon). Only one such possible topology is represented at the next step. The *bottom line* indicates the number of possibilities. These numbers multiply to obtain the total number of trees.

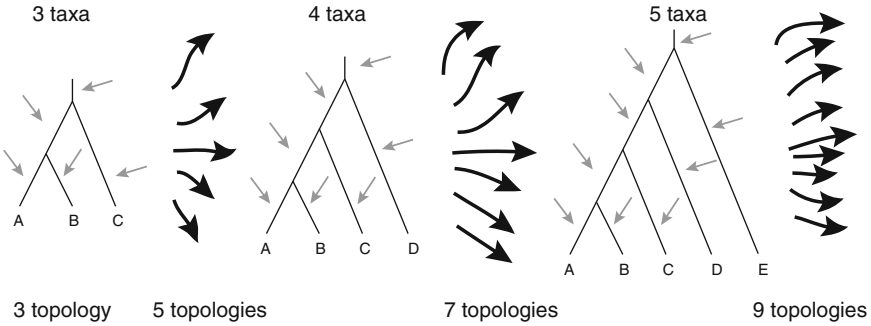

Fig. 10. Procedure to count the number of rooted topologies. See Fig. 9 for legend and text for details.

where the Γ function for any real number x is defined as $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$. An approximation based on Stirling number is also given in (84).

The same exercise can be done for rooted trees (Fig. 10), where the number of possible rooted topologies with n taxa becomes $1 \times 3 \times \cdots \times 2n - 3$, which is:

$$N_{rooted}^{T(n)} = \frac{(2n-3)!}{2^{n-2}(n-2)!} = \frac{2^{n-1}\Gamma(n-(1/2))}{\sqrt{\pi}} \tag{27}$$

Note that $N_{unrooted}^{T(n)} = N_{rooted}^{T(n-1)}$, as Table 2 clearly suggests.

As a result, the number of possible topologies quickly becomes very large when the number n of sequences increases, even with a very modest n, so that heuristics become necessary to find the best scoring tree.

2.10.2. Some Heuristics to Find the Best Tree

The simplest approach builds upon the idea presented in Figs. 9 and 10. Stepwise addition, for instance, starts with three sequences drawn at random among the n sequences to be analyzed, and adds sequences one at a time, keeping only the tree that has the highest

Table 2
Counting tree topologies

Number of taxa	Unrooted tree	Rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
10	2,027,025	34,459,425
20	221,643,095,476, 699,771,875	8,200,794,532,637, 891,559,375

Number of tree topologies are given for the unrooted and rooted cases

score at each step (e.g., ref. 3). However, there is no guarantee that the final tree is the optimal tree (35). The idea behind branch-and-bound (85), refined in ref. 86, is to have a look-ahead routine that prevents entrapment in suboptimal trees. This routine sets a bound on the trees selected at each round of additions, such that only the trees that have a score at least as good as that of the trees obtained in the next round are kept in the search algorithm. Solutions found by the branch-and-bound algorithm are optimal, but computing time becomes quickly prohibitive with more than 20 sequences.

As a result, most tree-search algorithms will start with a quickly obtained tree, often reconstructed with an algorithm based on pairwise distances such as Neighbor-Joining (87) or a related approach (88, 89), and then alter the tree randomly until no further improvement is obtained or after a certain number of unsuccessful attempts is reached. Examples of such algorithms include Nearest Neighbor Interchange (NNI), Subtree Pruning and Regrafting (SPR), or Tree Bisection and Reconnection (TBR); see, e.g., ref. 3 for a full description. While the details are of little importance here, the critical point is the extent of topological rearrangement in each case. With NNI for instance, each rearrangement can give rise to two topologies. The result is that exploring the topology space is slow, especially in problems with large n. On the other hand, TBR has, among the three methods cited above, the largest number of neighbors. As a result, the topology space is explored quickly, but the optimal tree can be "missed" simply because a dramatic change is attempted, so that the computational cost increases. Alternatively, the chance of finding the optimal tree $\hat{\tau}$ when $\hat{\tau}$ is very different from the current tree is higher when the algorithm can

create some dramatic rearrangements. Some programs, such as PhyML ver. 3.0, now use a combination of NNI and SPR to address this issue (90). MCMC samplers that search the tree space implement somewhat similar tree-perturbation algorithms that are either "global" and modify the topology dramatically, or "local" (91) (see also ref. 92 for a correction of the original local moves). As a result, MCMC samplers are affected by the same issues as traditional likelihood methods. Much of the difficulty therefore comes from this kind of trade-off between larger rearrangements that are expected to improve accuracy and the computational burden associated with these extra computations (93).

3. Uncovering Processes and Times

3.1. Dating the Tree of Life: Always Deeper?

Similar to the problem of estimating the tree of life, dating the tree of life poses many challenges (94). Since it was first proposed in 1965 (32), the idea of estimating divergence times has since undergone a dramatic change, and new approaches are regularly proposed. Population geneticists have their own approaches, which are either fully Bayesian (95) or based on Approximate Bayesian Computation in the coalescent framework (96). All these approaches make it possible to infer divergence times between recently diverged species, as in the case of humans and chimpanzees, or to date demographic events such as the migrations "out of Africa" of early human populations (97).

In the context of molecular evolution, we are usually interested in estimating deeper divergence times, such as those between species, which are available online for instance at www.timetree.org (98) (check also the corresponding app for smartphones). While early "molecular dates" were systematically biased towards ages that are too old (94), we argue here that recent developments in the field have led to more accurate methods and also to a better understanding of methodological limitations.

3.1.1. The Strict Molecular Clock

One quantity that we can estimate when comparing pairs of sequences is the number of differences that exist. This number, estimated as a branch length b, can be corrected for multiple substitutions (see Subheading 2.7), but basically remains an expected number of substitutions per site. With "dating" (defined here as the activity of estimating divergence times (99)), we are interested in estimating time t, which relates to the expected numbers of substitutions b according to the following equation:

$$b = \Delta t \times r \tag{28}$$

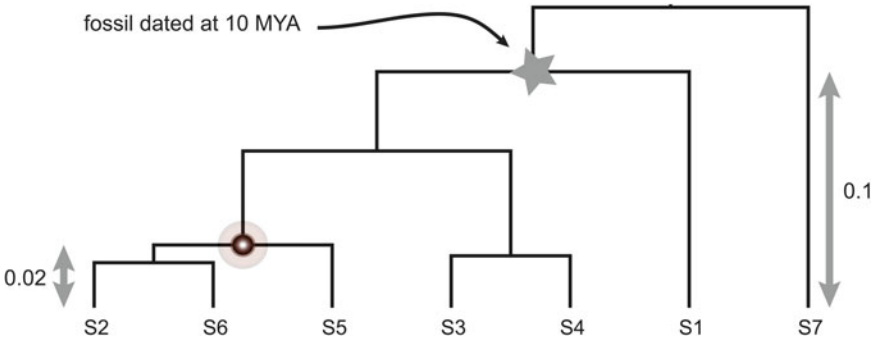

Fig. 11. The strict molecular clock. The tree is ultrametric. The node marked with a *star* indicates the presence of a fossil, dated in this example to ten MYA. This is the point that we will use to calibrate the clock, that is, to estimate the global rate of evolution. The number of substitutions from the marked node to the tips (present) is indicated on the right weights in at 0.1 substitutions/site. The node that is the most recent common ancestor of S2 and S5 is the node of interest. The number of substitutions from this node to the tips is 0.02 substitutions/site.

where Δt is a period of time and r the rate of evolution. In technical terms, times and rates are said to be confounded, because we cannot estimate one without making an assumption about the other.

The molecular clock hypothesis does just this by assuming that rates of evolution are constant in time (32) (see also ref. 100, p. 65). Under this assumption, the estimated tree is ultrametric as in the toy example represented in Fig. 11, which implies that all the tips are level, or equivalently that the distance from root to tip is the same for all branches.

In this example (Fig. 11), the branch length from the fossil-dated node is 0.1 substitutions/site (sub/site), and the fossil was estimated to be present ten million years ago (MYA). Under the strict molecular clock assumption (equal rates over the whole tree), we can (1) estimate the rate of evolution ($0.1/10 = 0.01$ sub/site/my) and (2) date all the other nodes on the tree. For instance, the most recent common ancestor of S2 and S5 is separated from the tips by a branch length of 0.02 sub/site. Its divergence time is therefore $0.02/0.01 = 2$ MYA.

As with any hypothesis, the strict clock can be tested. Tests based on relative rates assess whether two species evolve at the same rate as a third one, used as an outgroup. Originally formulated in a distance-based context (101), likelihood versions have been described (35, 102). However, because of their low power (103) their use is on the wane. The most powerful test is again the LRT (see Subheading 2.9). The test proceeds as usual, first calculating the test statistic $2\Delta\ell$ (twice the difference of log-likelihood values). The null hypothesis (strict clock) is nested within the alternative hypothesis (clock not enforced), so that $2\Delta\ell$ follows a χ^2 distribution. The degree of freedom is calculated following Fig. 12. With

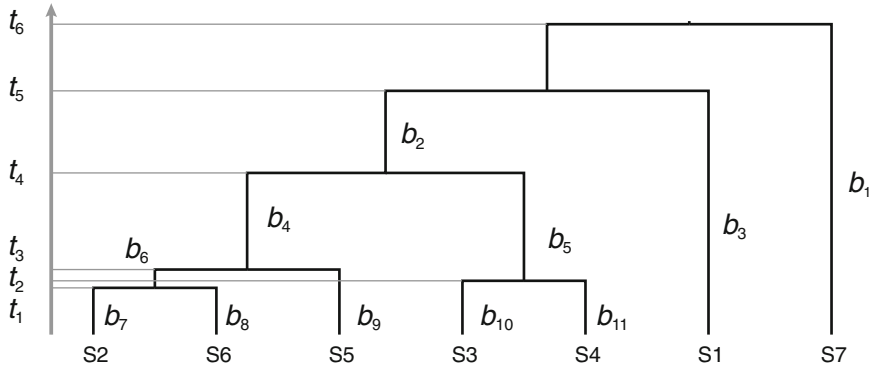

Fig. 12. Testing the strict molecular clock. The divergence times that can be estimated under the strict clock assumption are denoted t_i. The branch lengths that can be estimated without the clock are denoted b_i. In the case depicted, with $n = 7$ sequences, we have $n - 1 = 6$ divergence times and $2n - 3 = 11$ branch lengths.

an alignment of n sequences, we can estimate $n - 1$ divergence times under the null model (disregarding parameters of the substitution model) and we have $2n - 3$ branch lengths under the alternative model. The difference in number of free parameters is therefore $n - 2$, which is our degree of freedom. This version of the test actually assesses whether all tips are at the same distance from the root of the tree (35). For time-stamped data, serially sampled in time as in the case of viruses, the alternative model incorporates information on tip dates (104).

This linear regression model suggested by the molecular clock hypothesis has often been portrayed as a recipe (105), which gave rise in the late twentieth to early twenty-first century to a veritable cottage industry (106–109), culminating with a paper suggesting that the age of the tree of life might be older than the age of planet Earth (110). This recipe was put down by two factors: (1) the publication of a piece written in a rather unusual style for a scientific paper (111) and (2) new methodological developments. The main points made in (111) are that (1) most of the early dating studies relied on one analysis (107) that used a fossil-based calibration point for the divergence of birds at 310 MYA to estimate a number of molecular dates for vertebrates, and that (2) these molecular dates were then used in subsequent studies as a proxy for calibration points, disregarding their uncertainty. As a result, estimation errors were passed on and amplified from study to study, leading to the nonsensical results in (110).

3.1.2. Local Molecular Clocks

This "debacle" has motivated further theoretical developments in the dating field. The simplest idea is that, if a global clock does not hold for the entire tree, then perhaps groups of related species share the same rate. That is, if a *global* clock does not hold, perhaps the tree can be subdivided into *local* molecular clocks. An initial idea was proposed in the context of quartets of sequences (112) and was

later generalized to a tree of any size with any number of local clocks on the tree (113) (constrained by the number of branches on the tree and calibration points). Because of the arbitrariness of such local clocks, methods have been devised to place the clocks on the tree (114) and to estimate the appropriate number of clocks that should be used (115). A Bayesian approach now estimates all these parameters and their placement in an integrated statistical framework (116).

3.1.3. Correlated Relaxed Clocks

The idea of a correlated relaxed molecular clock goes back to Sanderson (117) (see also ref. 118), who considered that rates of evolution can change from branch to branch on a tree. By constraining rates of evolution to vary in an autocorrelated manner on a tree, it is possible to devise a method that minimizes the amount of rate change.

The idea of an autocorrelated process governing the evolution of the rates of evolution is attributed to ref. 119 in ref. 117, but could all the same be attributed to Darwin. Thorne, Kishino, and coworkers (120) developed this idea further in a Bayesian framework. Building upon the basic theory covered in Subheading 2.9, the idea is to place prior distributions on the quantities in the right-hand side of Eq. 28. The target distribution is $p(t|X)$. It is proportional to $p(X|t)p(t)$ according to Bayes' theorem, but all that we can estimate is:

$$p(b|X) = \frac{p(X|b)p(b)}{p(X)} = \frac{p(X|r,t)p(r,t)}{p(X)} \tag{29}$$

One way of expanding the joint distribution of rates and times is $p(r, t)$ is $p(r|t)p(t)$, which posits a process where rate change depends on the length of time separating two divergences. The "art" is now in choosing prior distributions, conditional on the obvious constraint that rates and times should take positive values. A number of such prior distributions for rates have been proposed and assessed (121) and one of the best performing model for rates is, in our experience, the log-normal model (120, 122). The prior on times is either a pure-birth (Yule) model or a birth-and-death process possibly incorporating species sampling effects (123). If sequences are sampled at the population level, a coalescent process is more appropriate (see ref. 124 for an introduction). In this case, the past demography of the sampled sequences can be traced back taking inspiration from spline regression techniques (125, 126) or multiple change-point models (127).

Once these priors are specified, an MCMC sampler will draw from the target distribution in Eq. 29, and marginal distributions for times and rates can easily be obtained. The rationale behind the sampler is represented in Fig. 13. As per Eq. 28, the relationship between rates and time is the branch of a hyperbolic curve, where

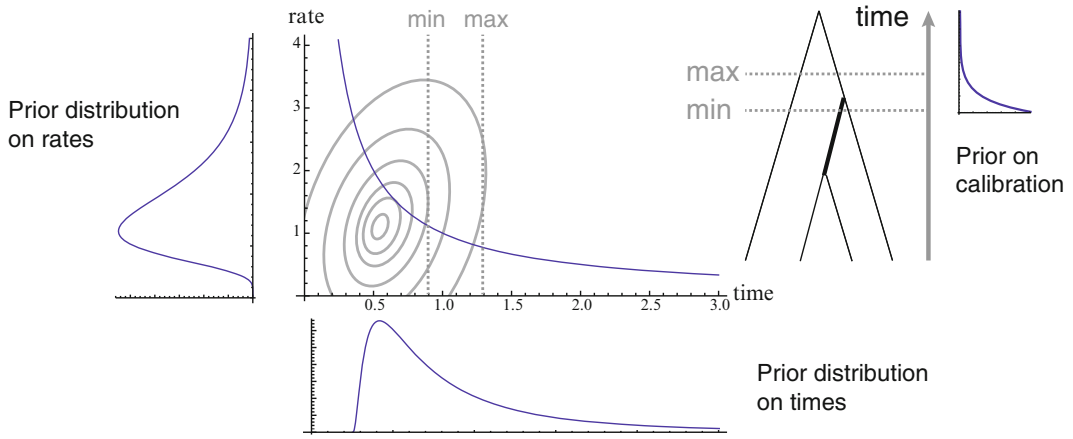

Fig. 13. The relaxed molecular clock. See text for details.

the priors on rates and on times define a region of higher posterior probability, symbolized here by a contour plot superimposed on the hyperbolic curve. On top of this, fossil information is incorporated into the analysis as constraints on times (111) stimulated a discussion about the shape of these prior distributions, which was taken up in ref. 128, and further developed in ref. 129. Briefly, fossil information is usually imprecise, as paleontologists can only provide minimum and maximum ages (Fig. 13). Of these two ages, the minimum age is often the most reliable. Under the assumption that the placement of the fossil on the tree is correct, the idea is to place on fossil dates a prior distribution that will be highly skewed towards older (maximum) ages. A "hard bound" can be placed on the minimum age, possibly by shifting this prior distribution by an offset equal to the minimum age, while the tails of the prior distribution will act as "soft bounds," because they do not impose on the tree a strict (or *hard*) constraint. Empirical studies agree, however, that both reliability and precision of fossil calibrations are critical to estimating divergence times (95, 130).

3.1.4. Uncorrelated Relaxed Clocks

Because of the autocorrelation between the rate of each branch and that of its ancestral branch (except for the root, which obviously requires a special treatment), the tree topology is fixed under the autocorrelated models described above. By relaxing this assumption about rate autocorrelation, Drummond et al. (131) were able to implement a model that also integrates over topological uncertainty. In spite of the somewhat counter-intuitive nature of the relaxation of the autocorrelated process, empirical studies have found this approach to be one of the best performing (e.g., ref. 115).

When first published, it was proposed that making use of an uncorrelated relaxed molecular clock could improve phylogenetic

inference (131). The idea was that calibration points and their placement on the tree could act as additional information. However, a simulation study suggests that relaxed molecular clocks might not improve phylogenetic accuracy (132), a result that might be due to the lack of calibration constraints in this particular simulation study.

3.1.5. Some Applications of Relaxed Clock Models

Since the advent of relaxed molecular clocks, two very exciting developments have seen the light of day. The first concerns the inclusion of spatial statistics into dating models (133, 134). Spatial statistics are not new in population genetics (135) and have been used with success in combination with analyses in computational molecular evolution (e.g., ref. 136). However, the originality in ref. 134 for instance is to combine in a single statistical framework molecular data with geographical and environmental information to infer the diffusion of sequences through both space and time. While these preliminary models seem to deal appropriately with natural barriers to gene flow such as coastlines, a more detailed set of constraints on gene flow may further enhance their current predictive power.

The second development coming from relaxed molecular clocks concerns the mapping of ancestral characters onto uncertain phylogenies. This is not a novel topic, as a Bayesian approach was first described in 2004 (137, 138). The novelty is that we now have the tools to correlate morphological and molecular evolution in terms of their absolute rates and to allow both molecular and morphological rates of evolution to vary in time (139). Further development will certainly integrate over topological uncertainty. While there has been a heated controversy about the existence of such a correlation in the past (140), all previous studies were using branch length as a proxy for rate of molecular evolution, which is clearly incorrect. We can therefore expect some more accurate results on this topic very soon.

4. Molecular Population Phylogenomics

Population genetics is rich in theory regarding the relative roles of mutation, drift, and selection. Much research in population genomics is now focusing on using this theory to develop statistical procedures to infer past processes based on population-level data, such as those of the 1,000-genome project (141). One limitation of these inference procedures is that they all focus on a thin slice of evolutionary time by studying evolution at the level of populations. If we wish to study longer evolutionary time scales, for example, tens or hundreds of millions of years, we must resort to interspecific data. In such a context, which is becoming intrinsically *phylogenetic*,

the most important event is a substitution, that is, a mutation that has been fixed. Yet substitution rates can be defined from several features. In particular, from a population genetics perspective, it is of interest to model both mutational features and selective effects, combining them multiplicatively to specify substitution rates. We review briefly how substitution models that invoke codons as the state space lend themselves naturally to these objectives in a first section below (Subheading 1), before explaining the origin (and a shortcoming) of all the approaches developed so far (Subheading 2).

4.1. Bridging the Gap Between Population Genetics and Phylogenetics

Assuming a point-mutation process, such that events only change one nucleotide of a codon during a small time interval, Muse and Gaut proposed a codon substitution model with rates specified from the Q_{GTR} nucleotide-level matrix (see Subheading 2.7), along with one parameter that modulates synonymous events and another one that modulates nonsynonymous events (142). In most subsequent formulations, the parameter associated with synonymous events is assumed to be fixed, such that the model only modulates nonsynonymous rates by means of a parameter denoted ω. This parameter has traditionally been interpreted as the nonsynonymous to synonymous rate ratio, and is generally associated with a different formulation of the codon model proposed by Goldman and Yang (143). More details on codon models can be found in Chapter 5 in Volume 2 (144). There continues to be a debate regarding the interpretation of the ω parameter (145, 146). Regardless of how this issue is settled, it is clear that ω is aimed at capturing the net overall effects of selection, irrespective of the exact nature of these effects.

With the intention to model selective effects themselves, Halpern and Bruno (147) proposed a codon substitution model that combines a nucleotide-level layer, as described above, for controlling mutational features, along with a fixation factor that is proportional to the fixation probability of the mutational event. The fixation factor is in turn specified from an account of amino acid or codon preferences. One objective of the model, then, consists in teasing apart mutation and selection. While in ref. 147 proposed their model with site-specific fixation factors, later work has explored simpler specifications, where all sites have the same fixation factor (148). Other models that aimed at capturing across-site heterogeneities in fixation factors were proposed using nonparametric devices and empirical mixtures (149). Recent developments include sequence-wide fixation factors (145, 150), and we predict that these models will play a role in bridging the gap between molecular evolution at the population and at the species levels.

4.2. Origin of Mutation-Selection Models: The Genic Selection Model

In order to understand a shortcoming of these models, we need to go back to the development of fixation probabilities that took place in the second half of the twentieth century. The basic unit or *quantum* of evolution is a change in allele frequency p. Allele frequencies can be affected by four processes: migration, mutation, selection, and drift. Because of the symmetry between migration and mutation (151), which only differ in their magnitude, these two processes can be treated as one. We are left with three forces: mutation, selection, and drift. The question is then, what is the fate of an allele under the combined action of these processes? Our development here follows (152) (but see ref. 153 for a very clear account).

4.3. Fixation Probabilities

Of the three processes affecting allele frequencies, mutation, and selection can be seen as directional forces in that their action will shift the distribution of allele frequencies towards a particular point, be it an internal equilibrium, or fixation/loss of an allele. On the other hand, drift is a nondirectional process that will increase the variance in allele frequencies across populations, and will therefore spread out the distribution of allele frequencies. This distribution is denoted $\Psi(p, t)$. We also must assume that the magnitude of all three processes, mutation, selection, and drift, is small and of the order of $1/(2N_e)$, where N_e is the effective population size. To derive the fate of an allele after a certain number of generations, we also need to define $g(p, \varepsilon; \mathrm{d}t)$, the probability that allele frequency changes from p to $p + \varepsilon$ during a time interval $\mathrm{d}t$.

In phylogenetics (and population genetics) we are generally interested in predicting the past. The tool making this possible is called the Kolmogorov backward equation, which predicts the frequency of an allele at some time t, given its frequency p_0 at time t_0:

$$\Psi(p,\ t+\mathrm{d}t|p_0) = \int \Psi(p,\ t|p_0\ +\varepsilon)g(p_0,\varepsilon;\mathrm{d}t)\ \mathrm{d}\varepsilon \qquad (30)$$

We can take the Taylor expansion of Eq. 30 around p_0, neglect all terms whose order is larger than two $\left(o\left(p_0^2\right)\right)$ and since Ψ is not a function of ϵ, we obtain:

$$\Psi(p,\ t+\mathrm{d}t|p_0) = \Psi\int g\ \mathrm{d}\varepsilon + \frac{\partial\Psi}{\partial p_0}\int \varepsilon g\ \mathrm{d}\varepsilon + \frac{\partial^2\Psi}{\partial p_0^2}\int \frac{\varepsilon^2}{2} g\ \mathrm{d}\varepsilon \qquad (31)$$

This formulation leads to the definition of two terms that represent the directional processes affecting allele frequencies (M) and the nondirectional process, or drift (V):

$$\begin{cases} M(p)\ \mathrm{d}t &= \int g\ \varepsilon\ \mathrm{d}\varepsilon \\ V(p)\ \mathrm{d}t &= \int g\ \varepsilon^2 \mathrm{d}\varepsilon \end{cases} \qquad (32)$$

Table 3
The standard selection models

Selection coefficients	A_1A_1	A_1A_2	A_2A_2
Genic (positive) selection	$w_1 = 1 + s$	$w_2 = 1 + hs$	$w_3 = 1$
Overdominance	$w_1 = 1$	$w_2 = 1 + s$	$w_3 = 1$

Models are represented for one locus with two alleles, A_1 and A_2, which define three genotypes A_1A_1, A_1A_2, and A_2A_2 of fitness w_1, w_2, and w_3. The selection coefficient is s (positive in this table, but not necessarily so) and the dominance is governed by h ($h \in [0, 1]$)

that we can substitute into Eq. 31. At equilibrium, $(\partial\Psi)/(\partial t) = 0$ and, after a bit of calculus, we obtain:

$$\frac{\partial\hat{\Psi}}{\partial p_0} = C\, e^{-\int [(2M)/(V)]\mathrm{d}p} \tag{33}$$

for which we need to specify boundary conditions and a model of selection. The boundary conditions are the two absorbing states of the system: (1) once fixed, an allele remains fixed ($\Psi(1, \infty; 1) = 1$) and (2) once lost, an allele remains lost ($\Psi(1, \infty; 0) = 0$). With these two requirements, the probability that the allele frequency is 1 given that it was p_0 in the distant past is the fixation probability:

$$\Psi(1, \infty; p_0) = \frac{\int_0^{p_0} e^{-\int [(2M)/(V)]\mathrm{d}p}\mathrm{d}p}{\int_0^1 e^{-\int [(2M)/(V)]\mathrm{d}p}\mathrm{d}p} \tag{34}$$

We therefore only need to compute M and V under a particular model of selection to fully specify the fixation probability of an allele in a mutation–selection–drift system. All that is required now to go further is a selection model.

4.4. The Case of Genic Selection

We are now ready to derive an explicit form to $\Psi(1, \infty; p_0)$ in Eq. 34 in the case of the genic selection model (Table 3; ref. 154). We obtain:

$$\bar{w} = 1 + sp^2 + 2pqhs = 1 + 2phs + sp^2(1 - 2h) \tag{35}$$

which can be approximated by $1 + 2phs$ (the result is exact only when $h = 1/2$). Therefore, $\mathrm{d}\bar{w}/\mathrm{d}p = 2hs$, and we can calculate the M and V terms to obtain the popular result:

$$\Psi(1, \infty; p_0) = \frac{\int_0^{p_0} e^{-\int [(2M)/(V)]\mathrm{d}p}\mathrm{d}p}{\int_0^1 e^{-\int [(2M)/(V)]\mathrm{d}p}\mathrm{d}p} = \frac{e^{-4N_e hsp_0} - 1}{e^{-4N_e hs} - 1}. \tag{36}$$

Now, the initial frequency of a mutation in a diploid population of (census) size N is $p_0 = 1/(2N)$ (following refs. 153; ref. 152

considered that $p_0 = 1/(2N_e)$; this debate is beyond the scope of this chapter), which leads to:

$$\Psi\left(1,\ \infty;\frac{1}{2N}\right) = \frac{e^{-2N_e hs/N} - 1}{e^{-2N_e hs} - 1} \tag{37}$$

If N_e is of the order of N, the numerator of the right-hand side of Eq. 37 becomes approximately $e^{-2hs} - 1$, whose Taylor approximation around $hs = 0$ is simply $-2hs$. We then obtain the result used in ref. 147, and in all the papers that implemented mutation–selection (–drift) models (e.g., refs. 145, 147–150):

$$\Psi\left(1,\infty;\frac{1}{2N}\right) = \frac{2hs}{1 - e^{-4N_e hs}} \tag{38}$$

Two critical points should be noted here. First, none of the recent codon models (145, 147–150) ever investigated the role of dominance h, as they all consider that the allele under (positive) selection is fully dominant. Second, Table 3 shows that another class of selection models, those based on balancing selection, has never been considered so far. The impact of the selection model on the predictions made by the mutation–selection (–drift) models is currently unknown.

5. High-Performance Computing for Phylogenetics

5.1. Parallelization

Because of the dependency of the likelihood computations on the shape of a particular tree (see Subheading 2.6), most phylogenetic computations cannot be parallelized to take advantage of a multiprocessor (or multicore) environment. Nevertheless, two main directions have been explored to speed up computations: first, in computing the likelihood of substitution models that incorporate among-site rate variation and second, in distributing bootstrap replicates to several processors, as both types of computations can be done independently. A third route is explored in Chapter 22 of Volume (155).

In the first case, among-site rate variation is usually modeled with a Γ distribution (156) that is discretized over a finite (and small) number of categories (157). The likelihood then takes the form of a weighted sum of likelihood functions, one for each discrete rate category, so that each of these functions can be evaluated independently. The route most commonly used is the plain "embarrassingly parallel" solution, where completely independent computations are farmed out to different processors. Such is the case for bootstrap replicates, for which a version of PhyML (90) exists, or in a Bayesian context for independent MCMC samplers (158) (see Subheading 2.9).

5.2. HPC and Cloud-Computing

More recent work has focused on the development of heuristics that make large-scale phylogenetics amenable to high-performance computing (HPC), that are performed on computer clusters. Because of the algorithmic complexity of resolving phylogenetic trees, an approach based on "algorithmic engineering" was developed (159). The underlying idea is akin to the training phase in supervised machine learning (160), except that here the target is not the performance of a classifier but that of search heuristics. All of these heuristics reuse parameter estimates, avoid the computation of the full likelihood function for all the bootstrap replicates, or seed the search algorithm for every n replicate on the results of previous replicates (159). For instance, in the "Lazy Subtree Rearrangement" (161), topologies are modified by SPR (see Subheading 2.10), but instead of recomputing the likelihood on the whole tree, only the branch lengths around the perturbation are reoptimized. This approximation is used to rank candidate topologies, and the actual likelihood is evaluated on the complete tree only for the best candidates. These heuristics now permit the analysis of thousands of sequences in a probabilistic framework (162), but the actual convergence of these algorithms remains difficult to evaluate, especially on very large data sets (e.g., $>10^4$ sequences).

In addition to the reduction of the memory footprint (163) in the case of sparse data matrices, an alternative direction to "tweaking likelihood algorithms" has been to take direct advantage of the computing architecture available. One particular effort aims at tapping directly into the computing power of graphics processing units or GPUs, taking advantage of their shared common memory, their highly parallelized architecture and the comparatively negligible cost of spawning and destroying threads on them. As a result, it is possible to distribute some of the summation entering the pruning algorithm (see Subheading 2.6) to different GPUs (164). The number of programs taking advantage of these developments is still limited to BEAST (78), mostly because CUDA (Compute Unified Device Architecture, up to version 1.3), the computing engine of these cards, was not IEEE-754 compliant and prone to numerical errors on double-precision computations. However, we anticipate that further programs will take advantage of GPUs as soon as newer cards fully support double-precision computation.

All these fast algorithms can be installed either on a local computer cluster, a solution adopted by many research groups in the recent past. However, installing a cluster can be demanding and costly because a dedicated room is required with appropriate cooling and power supply (not to mention securing the room, physically). Besides, redundancy requirements, both in terms of power supply and data, may demand hiring a system administrator. An alternative is to run analyses on a remote HPC server. Canada, for instance, has a number of such facilities thanks to national funding bodies (HPCVL at www.hpcvl.org, SHARCNET at www.sharcnet.ca, or HQCHP at

rqchp.ca, just to cite a few), and commercial solutions are just a few clicks away (e.g., Amazon Elastic Compute Cloud or EC2). Researchers can obtain access to these HPC solutions on a fee basis, either on demand or by means of a yearly subscription. But in spite of the technical support offered in the price, users still have to install their preferred phylogenetic software manually or put a formal request to the team of system administrators managing the HPC facility, all of which is not always convenient.

To make the algorithmic and technological developments described above more accessible, the recent past has seen the emergence of cloud-computing (165) dedicated to the phylogenetics community. Examples include iPlant (iplantcollaborative.org), CIPRESS (www.phylo.org), or Phylogeny.fr (www.phylogeny.fr (166)). Many include web portals that do not require that users be well versed in unix commands, and some portals such as iPlant plan to offer an application programming interface to cater to the most computer-savvy users. One potential limitation of these services is the bandwidth necessary to transfer large files. In our experience with MCMC samplers, the output of a single run can reach a half dozen gigabytes. Being text files, these output files can easily be compressed by an order of magnitude. The management of relatively large files will remain a potential issue, unless phylogenetics practitioners are ready to discard these files after analysis, the end product of which is a single tree file a few kilobytes in size, in the same way that people involved in genome projects delete the original image files produced by massively parallel sequencers. Data security or privacy might not be a problem in most applications, except in projects dealing with human subjects or viruses such as HIV that expose the sexual practices of subjects. However, once these various hurdles are out of the way, users could very well imagine running their phylogenetic analyses with millions of sequences from a smartphone app while commuting.

6. Conclusions

Although most of the initial applications of likelihood-based methods were motivated by the shortcomings of parsimony, they have now become well accepted as they constitute principled inference approaches that rely on probabilistic logic. Moreover, they allow biologists to evaluate more rigorously the relative importance of different aspects of evolution. The models presented in this chapter have the ability to disentangle rates from times (Subheading 3), or mutation from selection (Subheading 4), while in most cases accounting for the uncertainty about nuisance parameters. But the latest developments described above still make a number of restrictive assumptions (Subheading 2), and while many variations

in model formulations can be envisaged, they still remain to be explored in practice.

Although some progress has been made in developing integrative approaches (e.g., refs. 134, 139), throughout this chapter we assumed that a reliable alignment was available as a starting point. A number of methods exist to co-estimate an alignment and a phylogenetic tree (see Part II of this Volume), but the computational requirements and convergence of some of these approaches can be daunting, even on the smallest data sets by today's standards.

This brings us, finally, to the issue of tractability of most of these models in the face of very large data sets. The field of phylogenomics is developing quickly (see Part I of Volume 2), at a pace that is ever increasing given the output rate of whole genome sequencing projects. Environmental questions are drawing more and more attention, and metagenomes (see Part IV of Volume 2) will be analyzed in the context of what will soon be called *metaphylogenomics*. Exploring the numerous available and foreseeable substitution models in such contexts will require continued work in computational methodologies. As such, modeling efforts will continue to go hand-in-hand with, and maybe *dependent on*, algorithmic developments (167).

Acknowledgments

We would like to thank Michelle Brazeau, Eric Chen, Ilya Hekimi, Benoît Pagé, and, in particular, Wayne Sawtell for their critical reading of a draft of this chapter. This work was partly supported by the Natural Sciences Research Council of Canada (N.R., S.A.B.) and the University of Ottawa (S.A.B.).

References

1. Nei, M. and Kumar, S. (2000) *Molecular evolution and phylogenetics*. Oxford University Press, Oxford, UK.
2. Higgs, P. G. and Attwood, T. K. (2005) *Bioinformatics and molecular evolution*. Blackwell Pub, Malden, MA.
3. Yang, Z. (2006) *Computational molecular evolution*. Oxford University Press, Oxford, UK.
4. Balding, D. J., Bishop, M. J., and Cannings, C. (2007) *Handbook of statistical genetics*. John Wiley & Sons, 3rd ed edn, Chichester, UK.
5. Salemi, M., Vandamme, A.-M., and Lemey, P. (2009) *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press, 2nd edn, Cambridge, UK.
6. Aris-Brosou, S. and Xia, X. (2008) Phylogenetic analyses: A toolbox expanding towards Bayesian methods. *Int J Plant Genomics*, **2008**, 683509.
7. Rodrigue, N. and Philippe, H. (2010) Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends Genet*, **26**, 248–52.
8. Yang, Z. (2000) Complexity of the simplest phylogenetic estimation problem. *Proc Biol Sci*, **267**, 109–16.
9. Sober, E. (1988) *Reconstructing the past: parsimony, evolution, and inference*. MIT Press, Cambridge, MA.
10. Durbin, R. (1998) *Biological sequence analysis: probabalistic models of proteins and*

nucleic acids. Cambridge University Press, Cambridge, UK.

11. Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer Associates, Sunderland, MA.
12. Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**, 1586–91.
13. Efron, B. and Tibshirani, R. (1993) *An introduction to the bootstrap*, vol. 57. Chapman & Hall, New York, NY.
14. Efron, B., Halloran, E., and Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*, **93**, 7085–90.
15. Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, **39**, 783–791.
16. Baldauf, S. L. (2003) Phylogeny for the faint of heart: a tutorial. *Trends Genet*, **19**, 345–51.
17. Hasegawa, M. and Kishino, H. (1989) Confidence limits of the maximum-likelihood estimate of the hominoid three from mitochondrial-DNA sequences. *Evolution*, **43**, 672–677.
18. Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*, **55**, 539–52.
19. Hillis, D. M. and Bull, J. J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol*, **42**, pp. 182–192.
20. Felsenstein, J. and Kishino, H. (1993) Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. *Syst Biol*, **42**, pp. 193–200.
21. Yang, Z. and Rannala, B. (2005) Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst Biol*, **54**, 455–70.
22. Berry, V. and Gascuel, O. (1996) On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. *Mol Biol Evol*, **13**, 999.
23. Shimodaira, H. and Hasegawa, M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, **17**, 1246–7.
24. Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*, **27**, 401–410.
25. Tuffley, C. and Steel, M. (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull Math Biol*, **59**, 581–607.
26. Steel, M. and Penny, D. (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol Biol Evol*, **17**, 839–50.
27. Holder, M. T., Lewis, P. O., and Swofford, D. L. (2010) The Akaike Information Criterion will not choose the no common mechanism model. *Syst Biol*, **59**, 477–85.
28. Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005) Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol*, **5**, 50.
29. Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., and Philippe, H. (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol*, **54**, 743–57.
30. Hampl, V., Hug, L., Leigh, J. W., Dacks, J. B., Lang, B. F., Simpson, A. G. B., and Roger, A. J. (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups." *Proc Natl Acad Sci U S A*, **106**, 3859–64.
31. Liu, H., Aris-Brosou, S., Probert, I., and de Vargas, C. (2010) A timeline of the environmental genetics of the haptophytes. *Mol Biol Evol*, **27**, 161–76.
32. Zuckerkandl, E. and Pauling, L. (1965) Evolutionary divergence and convergence in proteins. Bryson, V. and Vogel, H. J. (eds.), *Evolving Genes and Proteins*, pp. 97–166, Academic Press, New York, NY.
33. Galtier, N., Gascuel, O., and Jean-Marie, A. (2005) Markov models in molecular evolution. Nielsen, R. (ed.), *Statistical Methods in Molecular Evolution*, pp. 3–24, Statistics for Biology and Health, Springer, New York, NY.
34. Cox, D. R. and Miller, H. D. (1965) *The theory of stochastic processes*. Wiley, New York, NY.
35. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, **17**, 368–76.
36. Jukes, J. C. and Cantor, C. R. (1969) Evolution of protein molecules. Munro, H. N. (ed.), *Mammalian protein metabolism*, pp. 21–123, Academic Press, New York, NY.
37. Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, **16**, 111–20.
38. Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, **22**, 160–74.
39. Tavaré, S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

40. Huelsenbeck, J. P., Larget, B., and Alfaro, M. E. (2004) Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol Biol Evol*, **21**, 1123–33.
41. Yang, Z. and Roberts, D. (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol Biol Evol*, **12**, 451–8.
42. Huelsenbeck, J. P., Bollback, J. P., and Levine, A. M. (2002) Inferring the root of a phylogenetic tree. *Syst Biol*, **51**, 32–43.
43. Yang, Z. (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Evol*, **51**, 423–32.
44. Aris-Brosou, S. (2005) Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol Biol Evol*, **22**, 200–9.
45. Burnham, K. P. and Anderson, D. R. (1998) *Model selection and inference: a practical information-theoretic approach*. Springer, New York, NY.
46. Anisimova, M., Bielawski, J. P., and Yang, Z. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*, **18**, 1585–92.
47. Whelan, S. and Goldman, N. (2004) Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, **167**, 2027–43.
48. Wong, W. S. W., Yang, Z., Goldman, N., and Nielsen, R. (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, **168**, 1041–51.
49. Massingham, T. and Goldman, N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–62.
50. Zhang, J., Nielsen, R., and Yang, Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, **22**, 2472–9.
51. Anisimova, M. and Yang, Z. (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol*, **24**, 1219–28.
52. Yang, Z. (2010) A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biol Evol*, **2**, 200–11.
53. Fletcher, W. and Yang, Z. (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol*, **27**, 2257–67.
54. Yang, Z. and dos Reis, M. (2011) Statistical properties of the branch-site test of positive selection. *Mol Biol Evol*, **28**, 1217–28.
55. Self, S. G. and Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *JASA*, **82**, 605–610.
56. Posada, D. and Crandall, K. A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–8.
57. Posada, D. (2008) jModelTest: phylogenetic model averaging. *Mol Biol Evol*, **25**, 1253–6.
58. Cunningham, C. W., Zhu, H., and Hillis, D. M. (1998) Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. *Evolution*, **52**, 978–987.
59. Pol, D. (2004) Empirical problems of the hierarchical likelihood ratio test for model selection. *Syst Biol*, **53**, 949–62.
60. Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *Ann Math Stat*, **22**, 79–86.
61. Minin, V., Abdo, Z., Joyce, P., and Sullivan, J. (2003) Performance-based selection of likelihood models for phylogeny estimation. *Syst Biol*, **52**, 674–83.
62. Ripplinger, J. and Sullivan, J. (2008) Does choice in model selection affect maximum likelihood analysis? *Syst Biol*, **57**, 76–85.
63. Posada, D. and Crandall, K. A. (2001) Selecting the best-fit model of nucleotide substitution. *Syst Biol*, **50**, 580–601.
64. Abdo, Z., Minin, V. N., Joyce, P., and Sullivan, J. (2005) Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. *Mol Biol Evol*, **22**, 691–703.
65. Luo, A., Qiao, H., Zhang, Y., Shi, W., Ho, S. Y., Xu, W., Zhang, A., and Zhu, C. (2010) Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. *BMC Evol Biol*, **10**, 242.
66. Schwarz, G. (1978) Estimating the dimension of a model. *Ann Stat*, **6**, 461–464.
67. Evans, J. and Sullivan, J. (2011) Approximating model probabilities in Bayesian Information Criterion and Decision-Theoretic approaches to model selection in phylogenetics. *Mol Biol Evol*, **28**, 343–9.
68. Kleinman, C. L., Rodrigue, N., Bonnard, C., Philippe, H., and Lartillot, N. (2006) A maximum likelihood framework for protein design. *BMC Bioinformatics*, **7**, 326.

69. Rodrigue, N., Philippe, H., and Lartillot, N. (2007) Exploring fast computational strategies for probabilistic phylogenetic analysis. *Syst Biol*, **56**, 711–26.
70. Yang, Z. (2005) Bayesian inference in molecular phylogenetics. Gascuel, O. (ed.), *Mathematics of Evolution and Phylogeny*, Chap. 3, pp. 63–90, Oxford University Press, Oxford, UK.
71. Jeffreys, H. (1939) *Theory of probability*. The International series of monographs on physics, The Clarendon press, Oxford, UK.
72. Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *JASA*, **90**, 773–795.
73. Lartillot, N. and Philippe, H. (2006) Computing Bayes factors using thermodynamic integration. *Syst Biol*, **55**, 195–207.
74. Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2011) Choosing among partition models in Bayesian phylogenetics. *Mol Biol Evol*, **28**, 523–32.
75. Newton, M. A. and Raftery, A. E. (1994) Approximating Bayesian inference with the weighted likelihood bootstrap. *J R Stat Soc B*, **56**, 3–48.
76. Aris-Brosou, S. (2003) How Bayes tests of molecular phylogenies compare with frequentist approaches. *Bioinformatics*, **19**, 618–24.
77. Ronquist, F. and Huelsenbeck, J. P. (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–4.
78. Drummond, A. J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, **7**, 214.
79. Raftery, A. E. (1996) Hypothesis testing and model selection. Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.), *Markov chain Monte Carlo in practice*, pp. 163–187, Chapman & Hall, Boca Raton, FL.
80. Xie, W., Lewis, P., Fan, Y., Kuo, L., and Chen, M.-H. (2011) Improving marginal likelihood estimation for Baysian phylogenetic model selection. *Syst. Biol*, **60**, 150–60.
81. Smyth, P. (2000) Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, **10**, 63–72.
82. Lartillot, N., Brinkmann, H., and Philippe, H. (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*, **7 Suppl 1**, S4.
83. Cavalli-Sforza, L. L. and Edwards, A. W. (1967) Phylogenetic analysis. models and estimation procedures. *Am J Hum Genet*, **19**, 233–57.
84. Aris-Brosou, S. (2003) Least and most powerful phylogenetic tests to elucidate the origin of the seed plants in the presence of conflicting signals under misspecified models. *Syst Biol*, **52**, 781–93.
85. Foulds, L. R., Penny, D., and Hendy, M. D. (1979) A general approach to proving the minimality of phylogenetic trees illustrated by an example with a set of 23 vertebrates. *J Mol Evol*, **13**, 151–166.
86. Hendy, M. D. and Penny, D. (1982) Branch and bound algorithms to determine minimal evolutionary trees. *Math Biosci*, **59**, 277–290.
87. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406–25.
88. Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*, **14**, 685–95.
89. Bruno, W. J., Socci, N. D., and Halpern, A. L. (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol*, **17**, 189–97.
90. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, **59**, 307–21.
91. Larget, B. and Simon, D. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol Biol Evol*, **16**, 750.
92. Holder, M. T., Lewis, P. O., Swofford, D. L., and Larget, B. (2005) Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. *Syst Biol*, **54**, 961–5.
93. Whelan, S. (2007) New approaches to phylogenetic tree search and their application to large numbers of protein alignments. *Syst Biol*, **56**, 727–40.
94. Benton, M. J. and Ayala, F. J. (2003) Dating the tree of life. *Science*, **300**, 1698–700.
95. Rannala, B. and Yang, Z. (2007) Inferring speciation times under an episodic molecular clock. *Syst Biol*, **56**, 453–66.
96. Wegmann, D., Leuenberger, C., and Excoffier, L. (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–18.
97. Reich, D., et al. (2010) Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, **468**, 1053–60.

98. Hedges, S. B., Dudley, J., and Kumar, S. (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**, 2971–2.
99. Welch, J. J. and Bromham, L. (2005) Molecular dating when rates vary. *Trends Ecol Evol*, **20**, 320–7.
100. Kimura, M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, UK.
101. Sarich, V. M. and Wilson, A. C. (1973) Generation time and genomic evolution in primates. *Science*, **179**, 1144–7.
102. Muse, S. V. and Weir, B. S. (1992) Testing for equality of evolutionary rates. *Genetics*, **132**, 269–76.
103. Bromham, L., Penny, D., Rambaut, A., and Hendy, M. D. (2000) The power of relative rates tests depends on the data. *J Mol Evol*, **50**, 296–301.
104. Rambaut, A. (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, **16**, 395–9.
105. Martin, A. P. (2001) Molecular clocks. *Encyclopedia of Life Sciences*, pp. 1–6, Nature Pub Group, New York, NY.
106. Wray, G. A., Levinton, J. S., and Shapiro, L. H. (1996) Molecular evidence for deep Precambrian divergences among Metazoan phyla. *Science*, **274**, 568–573.
107. Kumar, S. and Hedges, S. B. (1998) A molecular timescale for vertebrate evolution. *Nature*, **392**, 917–20.
108. Wang, D. Y., Kumar, S., and Hedges, S. B. (1999) Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc Biol Sci*, **266**, 163–71.
109. Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L., and Hedges, S. B. (2001) Molecular evidence for the early colonization of land by fungi and plants. *Science*, **293**, 1129–33.
110. Hedges, S. B., Chen, H., Kumar, S., Wang, D. Y., Thompson, A. S., and Watanabe, H. (2001) A genomic timescale for the origin of eukaryotes. *BMC Evol Biol*, **1**, 4.
111. Graur, D. and Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet*, **20**, 80–6.
112. Rambaut, A. and Bromham, L. (1998) Estimating divergence dates from molecular sequences. *Mol Biol Evol*, **15**, 442–8.
113. Yoder, A. D. and Yang, Z. (2000) Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol*, **17**, 1081–90.
114. Yang, Z. (2004) A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. *Acta Zool Sinica*, **50**, 645–56.
115. Aris-Brosou, S. (2007) Dating phylogenies with hybrid local molecular clocks. *PLoS One*, **2**, e879.
116. Drummond, A. J. and Suchard, M. A. (2010) Bayesian random local clocks, or one rate to rule them all. *BMC Biol*, **8**, 114.
117. Sanderson, M. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol*, **14**, 1218.
118. Sanderson, M. J. (2002) Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Mol Biol Evol*, **19**, 101–109.
119. Gillespie, J. H. (1991) *The causes of molecular evolution*. Oxford University Press, New York, NY.
120. Thorne, J. L., Kishino, H., and Painter, I. S. (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*, **15**, 1647–57.
121. Aris-Brosou, S. and Yang, Z. (2002) Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst Biol*, **51**, 703–14.
122. Aris-Brosou, S. and Yang, Z. (2003) Bayesian models of episodic evolution support a late precambrian explosive diversification of the Metazoa. *Mol Biol Evol*, **20**, 1947–54.
123. Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*, **43**, 304–11.
124. Hein, J., Schierup, M. H., and Wiuf, C. (2005) *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, Oxford, UK.
125. Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, **155**, 1429–37.
126. Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*, **22**, 1185–92.

127. Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol*, **25**, 1459–71.

128. Hedges, S. B. and Kumar, S. (2004) Precision of molecular time estimates. *Trends Genet*, **20**, 242–7.

129. Yang, Z. and Rannala, B. (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol*, **23**, 212–26.

130. Inoue, J., Donoghue, P. C. J., and Yang, Z. (2010) The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst Biol*, **59**, 74–89.

131. Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol*, **4**, e88.

132. Wertheim, J. O., Sanderson, M. J., Worobey, M., and Bjork, A. (2010) Relaxed molecular clocks, the bias-variance trade-off, and the quality of phylogenetic inference. *Syst Biol*, **59**, 1–8.

133. Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009) Bayesian phylogeography finds its roots. *PLoS Comput Biol*, **5**, e1000520.

134. Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010) Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol*, **27**, 1877–85.

135. Guillot, G., Santos, F., and Estoup, A. (2008) Analysing georeferenced population genetics data with geneland: a new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics*, **24**, 1406–7.

136. Nadin-Davis, S. A., Feng, Y., Mousse, D., Wandeler, A. I., and Aris-Brosou, S. (2010) Spatial and temporal dynamics of rabies virus variants in big brown bat populations across Canada: footprints of an emerging zoonosis. *Mol Ecol*, **19**, 2120–36.

137. Pagel, M. and Meade, A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol*, **53**, 571–81.

138. Pagel, M., Meade, A., and Barker, D. (2004) Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*, **53**, 673–84.

139. Lartillot, N. and Poujol, R. (2011) A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol*, **28**, 729–44.

140. Bromham, L., Woolfit, M., Lee, M. S. Y., and Rambaut, A. (2002) Testing the relationship between morphological and molecular rates of change along phylogenies. *Evolution*, **56**, 1921–30.

141. 1000 Genomes Project Consortium, Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–73.

142. Muse, S. V. and Gaut, B. S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, **11**, 715–24.

143. Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, **11**, 725–36.

144. Kosiol, C. and Anisimova, M. (2012) Methods for detecting natural selection in protein-coding genes. In Anisimova, M., (ed.), *Evolutionary genomics: statistical and computational methods* (volume 1). Methods in Molecular Biology, Springer Science+ Business media, LLC.

145. Thorne, J. L., Choi, S. C., Yu, J., Higgs, P. G., and Kishino, H. (2007) Population genetics without intraspecific data. *Mol Biol Evol*, **24**, 1667–77.

146. Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H., and Thorne, J. L. (2007) Quantifying the impact of protein tertiary structure on molecular evolution. *Mol Biol Evol*, **24**, 1769–82.

147. Halpern, A. L. and Bruno, W. J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*, **15**, 910–7.

148. Yang, Z. and Nielsen, R. (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, **25**, 568–79.

149. Rodrigue, N., Philippe, H., and Lartillot, N. (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A*, **107**, 4629–34.

150. Choi, S. C., Redelings, B. D., and Thorne, J. L. (2008) Basing population genetic inferences and models of molecular evolution upon desired stationary distributions of DNA or protein sequences. *Philos Trans R Soc Lond B Biol Sci*, **363**, 3931–9.

151. Hartl, D. L. and Clark, A. G. (2007) *Principles of population genetics*. Sinauer Associates, 4th ed edn, Sunderland, MA.

152. Kimura, M. (1962) On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713–9.

153. Rice, S. H. (2004) *Evolutionary theory: mathematical and conceptual foundations*. Sinauer Associates, Sunderland, MA.

154. Kimura, M. (1978) Change of gene frequencies by natural selection under population number regulation. *Proc Natl Acad Sci U S A*, **75**, 1934–7.

155. Prins, P., Belhachemi, D., Möller, S., and Smant, G. (2012) Scalable computing in evolutionary genomics. In Anisimova, M. (ed.), *Evolutionary genomics: statistical and computational methods* (volume 1). Methods in Molecular Biology, Springer Science+ Business media, LLC.

156. Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol*, **10**, 1396–401.

157. Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*, **39**, 306–14.

158. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, **20**, 407–15.

159. Stamatakis, A., Hoover, P., and Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol*, **57**, 758–71.

160. Hastie, T., Tibshirani, R., and Friedman, J. H. (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics, Springer, 2nd ed edn, New York, NY.

161. Stamatakis, A., Ludwig, T., and Meier, H. (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21**, 456–63.

162. Stamatakis, A., Göker, M., and Grimm, G. W. (2010) Maximum likelihood analyses of 3,490 rbcL sequences: Scalability of comprehensive inference versus group-specific taxon sampling. *Evol Bioinform Online*, **6**, 73–90.

163. Stamatakis, A. and Alachiotis, N. (2010) Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics*, **26**, i132–9.

164. Suchard, M. A. and Rambaut, A. (2009) Many-core algorithms for statistical phylogenetics. *Bioinformatics*, **25**, 1370–6.

165. Schatz, M. C., Langmead, B., and Salzberg, S. L. (2010) Cloud computing and the DNA data race. *Nat Biotechnol*, **28**, 691–3.

166. Dereeper, A., et al. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*, **36**, W465–9.

167. de Koning, A. P. J., Gu, W., and Pollock, D. D. (2010) Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. *Mol Biol Evol*, **27**, 249–65.

168. Anisimova, M. and Yang, Z. (2004) Molecular evolution of the hepatitis delta virus antigen gene: recombination or positive selection? *J Mol Evol*, **59**, 815–26.