PHYLOGENY ESTIMATION: TRADITIONAL AND BAYESIAN APPROACHES

Mark Holder and Paul O. Lewis

The construction of evolutionary trees is now a standard part of exploratory sequence analysis. Bayesian methods for estimating trees have recently been proposed as a faster method of incorporating the power of complex statistical models into the process. Researchers who rely on comparative analyses need to understand the theoretical and practical motivations that underlie these new techniques, and how they differ from previous methods. The ability of the new approaches to address previously intractable questions is making phylogenetic analysis an essential tool in an increasing number of areas of genetic research.

PHYLOGENETIC TREE
A graph depicting the ancestor–descendant relationships between organisms or gene sequences. The sequences are the tips of the tree. Branches of the tree connect the tips to their (unobservable) ancestral sequences.

SYSTEMATICS
The biological discipline that is devoted to characterizing the diversity of life and organizing our knowledge about this diversity (primarily through estimating the phylogenetic relationships between organisms).

Department of Ecology and Evolutionary Biology, 75 North Eagleville Road, University of Connecticut, Storrs, Connecticut 06269–3043, USA. Correspondence to M.H. email: mholder@uconn.edu
doi:10.1038/nrg1044

Comparative sequence analysis is an important tool for geneticists. Minutes after obtaining a new sequence, BLAST searches can give researchers hints about the function and other properties of a gene. Comparing several sequences along their entire length can show which parts are changing rapidly (and therefore might be less functionally constrained) and which residues show evidence of being shaped by natural selection[1]. Reconstructing ancestral sequences can show the timing and directionality of mutations[2]. These comparative analyses rely on a PHYLOGENETIC TREE that describes the evolutionary relationships between the sequences.

Estimating phylogenetic trees is not just an academic exercise: in some cases it can literally be a matter of life or death. For example, phylogenetic trees provided crucial evidence in the murder trial of a dentist that infected one of his patients with human immunodeficiency virus (HIV)[3]. Evolutionary trees also showed that cases of encephalitis in New York and New England represented the first examples of the mosquito-borne West Nile virus in the western hemisphere[4,5].

Any comparison of more than two related sequences implicitly assumes an underlying phylogeny. For some tasks, pairwise sequence comparisons are sufficient (for example, BLAST), but even these algorithms can be made more powerful by considering information from more than two sequences simultaneously (for example, PSI-BLAST). Given the potential power of comparative genomics (and its relatively low cost), even researchers with no interest in the evolution of organisms or genes *per se* should be aware of these approaches (BOX 1). Entire journals (such as *Systematic Biology* and *Molecular Phylogenetics and Evolution*) are largely devoted to estimating phylogenies. The myriad of conflicting approaches and the rich terminology of SYSTEMATICS can be intimidating to those who are interested in simply applying the methods. Nevertheless, the possibility of inaccurate tree estimates is real, so anyone who relies on comparative approaches should be familiar with the pitfalls in tree reconstruction.

Proceeding from the simple assumption that as the time increases since two sequences diverged from their last common ancestor, so does the number of differences between them, tree estimation seems to be a relatively simple exercise: count the number of differences between sequences and group those that are most similar. The simplicity of such an algorithm underestimates the complexity of the phylogenetic-inference problem. The rate of sequence evolution is not constant over time, so a simple measure of the genetic differences between sequences is not necessarily a reliable indication of when they diverged[6]. Natural selection or changing mutational biases during the history of an organism might cause

Box 1 | **Applications of phylogenetic methods**

Detection of orthology and paralogy
Phylogenetics is commonly used to sort out the history of gene duplications for gene families. This application is now included in even preliminary examinations of sequence data; for example, the initial analysis of the mouse genome[46] included neighbour-joining trees to identify duplications in cytochrome P450 and other gene families.

Estimating divergence times
Bayesian implementations of new models[37,38] allowed Aris-Brosou and Yang[40] to estimate when animal phyla diverged without assuming a molecular clock.

Reconstructing ancient proteins
Chang *et al.*[47] used maximum likelihood (ML) to reconstruct the sequence of visual pigments in the last common ancestor of birds and alligators; the protein was then synthesized in the laboratory (see REF. 48 for a recent discussion of the methodology of ancestral-character-state reconstruction).

Finding the residues that are important to natural selection
Amino-acid sites on the surface of influenza that are targeted by the immune system can be detected by an excess of non-synonymous substitutions[49–51]. This information might assist vaccine preparation.

Detecting recombination points
New Bayesian methods[52] can help determine which strains of human immunodeficiency virus-1 (HIV-1) arose from recombination.

Identifying mutations likely to be associated with disease
The lack of structural, biochemical and functional data from many genes implicated in disease means it is unclear which missense mutations are important. Fleming *et al.*[53] used Bayesian phylogenetics to identify missense mutations in conserved regions and regions under positive selection in the breast cancer gene *BRCA1*. These data allowed them to prioritize these mutations for future functional and population studies.

Determining the identity of new pathogens
Phylogenetic analysis is now routinely performed after polymerase chain reaction (PCR) amplification of genomic fragments of previously unknown pathogens. Such analyses made possible the rapid identification of both Hantavirus[54] and West Nile virus[4,5].

distantly related sequences to diverge from each other more slowly than is expected, or even become more similar to each other at some residues. Many of the sites in a DNA sequence are not helpful for phylogenetic reconstruction: some are functionally constrained so that they are invariant among all known sequences; others evolve rapidly (and, therefore, are not reliable indicators of deep relationships). As a result of such 'noise', often several phylogenetic hypotheses can explain the data reasonably well. So, the researcher must take this uncertainty into account. Commonly-used methods often ignore this uncertainty by producing a single estimate of the tree, which is then treated as the true tree. Additional (potentially time-consuming) steps are necessary to measure the strength of support for the groupings of sequences in the tree. Recently, BAYESIAN approaches to phylogenetic inference have provided a method for simultaneously estimating trees and obtaining measurements of uncertainty for every branch. Here, we briefly review the most popular methods of tree estimation (for a more thorough review, see REF. 6) and the newer Bayesian approaches.

Traditional approaches

The most common traditional approaches to reconstructing phylogenies are the neighbour-joining (NJ) algorithm and tree searches that use an optimality criterion such as PARSIMONY or maximum likelihood (ML). TABLE 1 shows a summary of the advantages and disadvantages of these methods, as well as a list of the software packages that implement them. BOX 2 sets these methods in the context of the entire study, from data collection to hypothesis testing.

Neighbour-joining. The extremely popular NJ algorithm[7,8] is relatively fast (a matter of seconds for most data sets) and, like all methods, performs well when the divergence between sequences is low. The first step in the algorithm is converting the DNA or protein sequences into a distance matrix that represents an estimate of the evolutionary distance between sequences (the evolutionary distance is the number of changes that have occurred along the branches between two

Table 1 | **Comparison of methods**

Method	Advantages	Disadvantages	Software
Neighbour joining	Fast	Information is lost in compressing sequences into distances; reliable estimates of pairwise distances can be hard to obtain for divergent sequences	PAUP* MEGA PHYLIP
Parsimony	Fast enough for the analysis of hundreds of sequences; robust if branches are short (closely related sequences or dense sampling)	Can perform poorly if there is substantial variation in branch lengths	PAUP* NONA MEGA PHYLIP
Minimum evolution	Uses models to correct for unseen changes	Distance corrections can break down when distances are large	PAUP* MEGA PHYLIP
Maximum likelihood	The likelihood fully captures what the data tell us about the phylogeny under a given model	Can be prohibitively slow (depending on the thoroughness of the search and access to computational resources)	PAUP* PAML PHYLIP
Bayesian	Has a strong connection to the maximum likelihood method; might be a faster way to assess support for treesthan maximum likelihood bootstrapping	The prior distributions for parameters must be specified; it can be difficult to determine whether the Markov chain Monte Carlo (MCMC) approximation has run for long enough	MrBayes BAMBE

For a more complete list of software implementations, see online link to Phylogeny Programs. For software URLs, see online links box.

BAYESIAN
A branch of statistics that focuses on the posterior probability of hypotheses. The posterior probability is proportional to the product of the prior probability and the likelihood.

PARSIMONY
In systematics, parsimony refers to choosing between trees on the basis of which one requires the fewest possible mutations to explain the data.

CONSENSUS METHOD
A summary of a set of trees in which branches that are not in most of the trees are collapsed to indicate uncertainty.

AGREEMENT SUBTREES
A tree containing the largest subset of sequences for which the relationships among sequences are invariant across all the phylogenies included.

sequences). A potentially serious weakness for distance methods such as NJ, is that the observed differences between sequences are not accurate reflections of the evolutionary distances between them. Multiple substitutions at the same site obscure the true distance and make sequences seem artificially close to each other. If the rate of evolution is constant over the entire tree (that is, if the 'molecular clock' hypothesis holds), correcting for multiple substitutions on tree estimation might be relatively unimportant[9].

Unfortunately, the molecular clock assumption is usually inappropriate for distantly related sequences. In this situation, a correction of the pairwise distances that accounts for multiple 'hits' to the same site should be used. There are many models for how sequence evolution occurs, each of which implies a different way to correct pairwise distances (see REF. 10 for some suggestions), so there is considerable debate on which correction to use (see BOX 2 for a discussion of model selection). Furthermore, these corrections have substantial variance when the distances are large.

In short, NJ is an extremely fast estimator of phylogenetic trees that does relatively well on clean data (for example, data from sequences that have diverged recently). When the goal is to infer older relationships, it can be difficult to arrive at reliable values for the distance matrix that is the input for NJ; obviously, if the input into an algorithm is poor, the algorithm has little chance of succeeding.

Tree searches under optimality criteria. The NJ tree is often treated as the starting point for a computationally intensive search for the best phylogeny (TABLE 2). To perform a tree search, a standard must be used for comparing trees — an optimality criterion, in the jargon of phylogenetics. The most popular criteria are parsimony, minimum evolution and ML. Minimum evolution (ME), like NJ, requires that the data be compressed into a distance matrix; therefore, like NJ, its success is dependent on the sequence divergences being adequately corrected for multiple hits (or the tree being so easy to infer that poor distance estimates are sufficient).

Box 2 | **The phylogenetic inference process**

The flowchart puts phylogenetic estimation (shown in the green box) into the context of an entire study. After new sequence data are collected, the first step is usually downloading other relevant sequences. Typically, a few outgroup sequences are included in a study to root the tree (that is, to indicate which nodes in the tree are the oldest), provide clues about the early ancestral sequences and improve the estimates of parameters in the model of evolution.

Insertions and deletions obscure which of the sites are homologous. Multiple-sequence alignment is the process of adding gaps to a matrix of data so that the nucleotides (or amino acids) in one column of the matrix are related to each other by descent from a common ancestral residue (a gap in a sequence indicates that the site has been lost in that species, or that a base was inserted at that position in some of the other species). Although models of sequence evolution that incorporate insertions and deletions have been proposed[55–58], most phylogenetic methods proceed using an aligned matrix as the input (see REF. 59 for a review of the interplay between alignment and tree inference).

In addition to the data, the scientist must choose a model of sequence evolution (even if this means just choosing a family of models and letting software infer the parameters of these models). Increasing model complexity improves the fit to the data but also increases variance in estimated parameters. Model selection[60–63] strategies attempt to find the appropriate level of complexity on the basis of the available data. Model complexity can often lead to computational intractability, so pragmatic concerns sometimes outweigh statistical ones (for example, NJ and parsimony are mainly justifiable by their speed).

As discussed in BOX 3, data and a model can be used to create a sample of trees through either Markov chain Monte Carlo (MCMC) or multiple tree searches on bootstrapped data (the 'traditional' approach). This collection of trees is often summarized using consensus-tree techniques, which show the parts of the tree that are found in most, or all, of the trees in a set. Although useful, CONSENSUS METHODS are just one way of summarizing the information in a group of trees. AGREEMENT SUBTREES are more resistant to 'rogue sequences' (one or a few sequences that are difficult to place on the tree); the presence of such sequences can make a consensus tree relatively unresolved, even when there is considerable agreement on the relationships between the other sequences. Sometimes, the bootstrap or MCMC sample might show substantial support for multiple trees that are not topologically similar. In such cases, presenting more than one tree (or more than one consensus of trees) might be the only way to appropriately summarize the data.

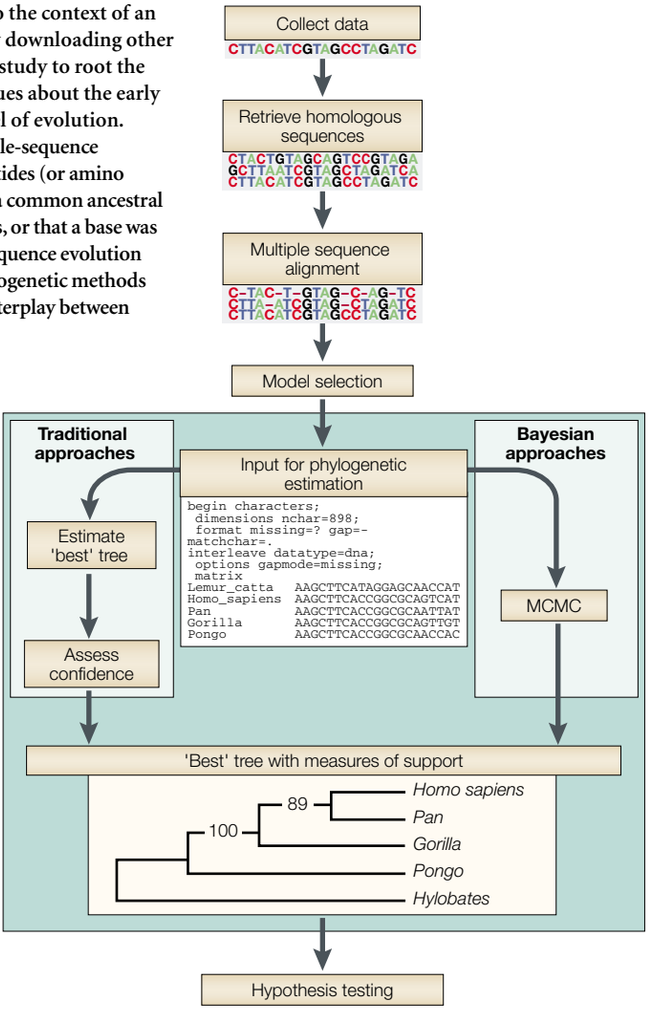


Table 2 | **Tree construction and tree searching methods**

Method	Description	Advantages	Disadvantages
Tree construction methods			
Stepwise addition	Builds a complete tree, starting with three sequences and attaching new sequences one at a time to the branch that yields the optimum tree at each step	Fast; later steps can reverse earlier pairing decisions	Yields one tree, often not global optimum; alternative additional sequences might yield different trees; not as fast as neighbour-joining
Star decomposition	Builds a completely resolved tree, starting with all sequences connected to a single 'hub' node. At each step, two lineages attached to the hub node are joined, becoming neighbours. Neighbours are chosen so that tree is optimal at each step	Fast; addition sequence irrelevant	Yields one tree, often not global optimum; neighbours cannot be dismantled at later steps; ties broken arbitrarily by some implementations
Neighbour joining	A star-decomposition method that uses an approximation to the minimum-evolution optimality criterion	One of the fastest of all tree construction methods	The same as those listed for star decomposition
Tree searching methods			
Heuristic search	Given a starting tree containing all sequences of interest, performs branch swapping to generate alternative trees in an attempt to find a better tree under a given optimality criterion. Strict hill-climber: if a better tree is found, the process begins again, stopping only if a local optimum is attained. Typically uses a stepwise addition or neighbour-joining tree as the starting tree	Faster than exact searches	Can miss the global optimal tree
Exact search	Exhaustive searches examine every possible tree and are guaranteed to return the best tree. Branch-and-bound techniques can eliminate some bad trees from consideration and still guarantee that they will return the best tree	The only methods that are guaranteed to find the best trees	Time-consuming: only practical for a few sequences (<20)

For every tree that is examined under ME, a set of branch lengths is estimated that comes closest to predicting the observed distances between sequences. The tree with the shortest size (the sum of the lengths of all of the branches) is judged to be the best estimate of the phylogeny. Some recent studies show that searching under ME is not much better than simply performing NJ[11], so the use of this criterion in computer-intensive tree searches might decline.

Parsimony. In contrast to distance-based approaches, parsimony and ML map the history of gene sequences onto a tree. By assessing the plausibility of the mutations that a particular tree would require to explain the data, a score can be assigned to each tree. In parsimony, the score is simply the minimum number of mutations that could possibly produce the data. There are fast algorithms that guarantee that any tree can be scored correctly[12–14].

Parsimony has a few obvious disadvantages. First, the score of a tree is completely determined by the minimum number of mutations among all of the reconstructions of ancestral sequences. Frequently, there are many plausible scenarios that could have produced a group of sequences. If one tree has few reasonable mutational pathways that could explain the data, whereas a second tree has many reconstructions with the same number of mutations, the second tree should be recognized as fitting the data better. An analogy would be tossing a pair of dice and predicting the sum. It would be a mistake to conclude that the numbers two and seven are equally probable outcomes from consideration of just the fact that the probability of seeing two ones, or a one and a six, are the same. The probability of observing a sum of seven is actually six times that of observing a sum of two, because there are six outcomes leading to a sum of seven, whereas only one outcome leads to a sum of two. In the same way, using just one mutational mapping as our guide might be misleading when the goal is determining which tree is the most plausible, because the mutational pathway along the evolutionary tree is unknown to us and we should consider all the possible paths that could explain the data (ML, which is discussed below, does consider all possible mutational mappings).

Another serious drawback of parsimony arises because it fails to account for the fact that the number of changes is unlikely to be equal on all branches in the tree. More changes would be expected to occur along the path linking primates to *Escherichia coli* than would be seen on the branches that link human to chimp. Unfortunately, when calculating the parsimony score of a tree, a mutation counts as one demerit to the score no matter where it occurs. Nucleotides that are present at the ends of long branches might be similar because of convergent evolution rather than direct inheritance, and parsimony does not allow for convergence along long branches as an explanation of similarity. This property makes parsimony susceptible to 'long-branch attraction'[15], in which two long branches that are not adjacent on the true tree are inferred to be the closest relatives of each other by parsimony.

Given that most phylogenies will have some long branches, the susceptibility of parsimony to long-branch attraction might seem damning. In reality, parsimony performs relatively well as long as the amount of convergence is rare compared with the number of mutations that are conveying useful information. Theory and simulation studies indicate that even divergent trees (in which there is substantial convergence) can be

inferred with a high degree of accuracy using parsimony, provided that the sequences are sampled densely enough[16,17]. The dense sampling of sequences means that the individual branches are short. Multiple changes at the same site rarely occur in one branch of such trees (it is also rare for the same site to be changed on a branch and one of its neighbouring branches). Under these conditions, the most parsimonious mapping is a good depiction of where changes occurred, and the parsimony scoring of trees works well. Achieving dense sampling of sequences is not always possible for practical or fundamental reasons (for example, because of extinction, the branch leading from the coelacanth to the other vertebrates will be a long branch that cannot be broken by better sampling).

Maximum likelihood. Accurately reconstructing the relationships between sequences that have been separated for a long time, or are evolving rapidly, requires a method that corrects for multiple mutational events at the same site. ML provides such a method. In ML, a hypothesis is judged by how well it predicts the observed data; the tree that has the highest probability of producing the observed sequences is preferred (this probability is the LIKELIHOOD of the tree). To use this approach, we must be able to calculate the probability of a data set given a phylogenetic tree. If we have a model of sequence evolution that describes the relative probability of various events (for example, the chance of a TRANSITION relative to the chance of a TRANSVERSION), then standard differential-equation techniques can be used to convert the model into a statement of the probability of any two sequences at opposite ends of a branch (for example, the probability that there will be an A on one end of a branch and a G on the other). These branch-based change probabilities take into account the possibility of unseen events (for example, back mutations or complex pathways such as A→C→G along the branch). To assess the probability of observed sequences on an entire tree requires considering the unobservable ancestral sequences. Ideally, every possible ancestral sequence would be considered when scoring a tree. Clever dynamic-programming tricks[18] make this feasible, although the calculation is much slower than scoring a tree under minimum evolution or parsimony.

From many perspectives, ML is the most appealing way to estimate phylogenies (see REF. 19 for a review of ML in phylogenetics). All possible mutational pathways that are compatible with the data are considered and the likelihood function is known to be a consistent and powerful basis for statistical inference in general[20]. The reliance on a model of sequence evolution might seem to restrict the use of ML to the realm of sequences that have an evolution that is already understood in detail. However, this is not the case. Relatively general models of evolution have been formulated. In these models, parameters govern aspects of the evolutionary process, such as the relative probabilities of transitions versus transversions, or the degree to which the rate of evolution differs across sites. The user does not need to know the correct values of these parameters; instead, they are estimated during the evaluation of trees. The value of a parameter that maximizes the likelihood is taken to be the appropriate value to use (see the section on joint versus marginal estimation below).

Although these general models are undoubtedly much simpler than the true process underlying sequence evolution, they seem to be relatively robust to violation of their simplifying assumptions. The main obstacle to the widespread use of ML is the computational burden. Algorithms that find the ML score must search through a multidimensional space of parameters. These techniques are not guaranteed to find the peak, but work relatively well for simple models of sequence evolution[21]. Unfortunately, they often have to score a tree hundreds of times, making large-scale problems (>100 sequences) tedious to solve with ML.

Assessing confidence — the bootstrap. A weakness of all of the methods discussed so far is that they produce only point estimates of the phylogeny; a computer program is run (sometimes for days) and the result is a tree (or group of trees that are equally good). The immediate question is 'how strongly does the data support each of the relationships depicted in the tree?'. Traditionally, this question has been tackled by a statistical technique called 'bootstrapping'[22,23] (although several other methods of assessing confidence exist[24]). In bootstrapping, the original data matrix is randomly re-sampled with replacement to produce pseudo-replicate data sets. The tree-building algorithm is performed on each of these replicate data sets. Bootstrapping offers a measure of which parts of the tree are weakly supported. A grouping that is present in a low percentage of the bootstrap replicates is sensitive to the exact combination of sites that were sequenced. This implies that if another data set were collected, there is a good chance that the group would not be recovered. Bootstrapping is a remarkably versatile tool (it can be used to assess the strength of support in virtually any type of analysis) that makes only minimal assumptions (although it does assume that each of the sites in the original data is independent of the others).

Comparative results that depend strongly on groups with low bootstrap support should be viewed with caution. The exact interpretation of the bootstrap proportion is elusive; higher is clearly better, but what is a reasonable cut-off? Some workers[25,26] have concluded that bootstrap proportions are conservative measures of support, so a value of 70% might indicate strong support for a group (for more discussion of the interpretation of bootstrap proportions, see REFS 27,28). It is also important to bear in mind that bootstrap proportions help predict whether the same result would be seen if more data were collected, not whether the result is correct. For example, in the case of a tree obtained because of long-branch attraction, bootstrapping might indicate relatively strong support[29] (because if you collected more data you would probably still be misled). So, bootstrapping cannot be

LIKELIHOOD
The probability of the data given the model and tree hypothesis. The likelihood measures how well the data agrees with the predictions made by the model and tree hypothesis.

TRANSITION
A mutation between two pyrimidines (T↔C) or two purines (A↔G).

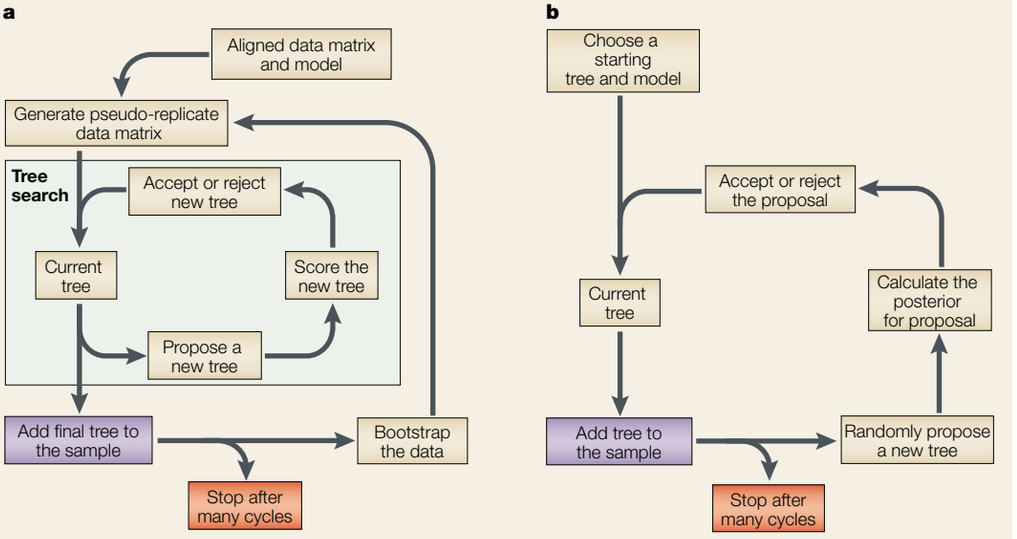
TRANSVERSION
A mutation between a pyrimidine and a purine (A↔C, A↔T, G↔C or G↔T).

Box 3 | **Bootstrapping and Markov chain Monte Carlo generate a sample of trees**

The number of times a particular group of sequences occurs in the trees from this sample can be used as a measure of how strongly the data supports that group. The bootstrapping approach (a) involves the generation of pseudo-replicate data sets by re-sampling with replacemtent the sites in the original data matrix. When optimality-criterion methods are used, a tree search (green box) is performed for each data set, and the resulting tree is added to the final collection of trees. A wide variety of tree-search strategies have been developed, but most are variants of the same basic strategy. An initial tree is chosen, either randomly or as the result of an algorithm — such as neighbour joining (NJ) or stepwise addition. Changes to this tree are proposed; the type of move can be selected randomly or the search can involve trying every possible variant of a particular type of move (TABLE 2; REF. 6). The new tree is scored and possibly accepted. Some search strategies are strict hill-climbers — they never accept moves that result in lower scores; others (genetic algorithms[64–66] or simulated annealing[67]) occasionally accept worse trees in an attempt to explore the tree space more fully. Making searching more accurate and faster is an active area of research[66,68]. For methods that use a tree-building algorithm, such as NJ, bootstrapping involves the application of the algorithm to each of the pseudo-replicate data sets instead of the tree-searching procedure.

The Markov chain Monte Carlo (MCMC) methodology (b) is similar to the tree-searching algorithm, but the rules are stricter. From an initial tree, a new tree is proposed. The moves that change the tree must involve a random choice that satisfies several conditions[43,44]. The MCMC algorithm also specifies the rules for when to accept or reject a tree.

Note that MCMC yields a much larger sample of trees in the same computational time, because it produces one tree for every proposal cycle versus one tree per tree search (which assesses numerous alternative trees) in the traditional approach. However, the sample of trees produced by MCMC is highly auto-correlated. As a result, millions of cycles through MCMC are usually required, whereas many fewer (of the order of 1,000) bootstrap replicates are sufficient for most problems.



used to overcome an inappropriate analysis of the data. It might be said that high bootstrap proportions are a necessary, but not sufficient, condition for having high confidence in a group.

The chief drawback of bootstrapping is the computational burden: the computational effort needed for the original analysis must be repeated several hundred times (once for each bootstrap replicate data set). This is not a concern when a fast analysis (like NJ) is employed, but it can be an obstacle when ML is used. BOX 3 summarizes the process of collecting a group of trees by bootstrapping.

Hypothesis testing. Bootstrapping gives coarse estimates of which parts of the tree are supported. Often, a researcher is interested in rejecting a specific hypothesis. Consider using a phylogenetic analysis to determine whether an unknown virus belongs to 'group A' or 'group B'. A tree with representatives of both candidate groups and the unknown sample is constructed, and the unknown sequence is intermingled with those from group A. Is it possible that the unknown sample was incorrectly placed because the data set is too small? After all, even if the data are equivocal, the unknown will be placed somewhere on the tree. The traditional approach to answering such a question involves finding the best tree in which the unknown sample clusters with the group B viruses, and then assessing how much worse this tree is compared to the best tree found in the original search. If the placement of the unknown with group B scores much worse than the optimal solution, then the data reject the possibility of

PRIOR PROBABILITY (The 'prior'). The probability of a hypothesis (or parameter value) without reference to the available data. Priors can be derived from first principles, or based on general knowledge or previous experiments.

the unknown sample actually belonging to group B. There are a number of statistical approaches[24] for assessing whether the difference in score between the preferred hypothesis and the null hypothesis is large enough to warrant rejecting the null. These techniques differ mainly in how they determine what constitutes a significantly different score — some rely on standard statistical distributions, whereas others use simulation techniques.

Bayesian phylogenetics

Bayesian approaches to phylogenetics are relatively new[30–35], but they are already generating a great deal of excitement because the primary analysis produces both a tree estimate and measures of uncertainty for the groups on the tree (TABLE 1; BOX 2). The field of Bayesian statistics is closely allied with ML. The optimal hypothesis is the one that maximizes the posterior probability. The posterior probability for a hypothesis is proportional to the likelihood multiplied by the PRIOR PROBABILITY of that hypothesis. Prior probabilities of different hypotheses convey the scientist's beliefs before having seen the data. In most applications, researchers specify prior probability distributions that they believe are largely uninformative, so that most of the differences in the posterior probability of hypotheses are attributable to differences in the likelihood. One way of doing this is to specify a uniform (or 'flat') prior in which every possible value of a parameter is given the same probability *a priori*. There is a considerable literature on the difficulties associated with finding priors that are uninformative (see REF. 36 for a statistical discussion); these issues can be serious, but are beyond the scope of this review.

In addition to providing measures of support faster than ML bootstrapping (BOX 3), Bayesian methods are exciting because they allow complex models of sequence evolution to be implemented. Three of the examples in BOX 1 (estimating divergence times, finding the residues that are important to natural selection and detecting recombination points) use Bayesian approaches to achieve aims that might be difficult or intractable for ML.

Complex parameter-rich models present two problems for ML. When the ratio of data points to parameters is low, ML estimates of parameters can be unreliable. In Bayesian analysis, the final result does not depend on one specific value, but considers all possible parameter values (see the section on joint estimation versus marginal estimation below). Even if there is enough data to estimate many parameters, the hill-climbing algorithms that are used to find the ML point can be slow or unreliable as the number of parameters increases (particularly if there are complex interactions between some of the parameters). Bayesian methods rely on an algorithm (see the Markov chain Monte Carlo (MCMC) section below) that does not attempt to find the highest point in the space of all parameters. The next section describes one example of how a Bayesian approach has given biologists a new way to study evolution.

Bayesian estimation of divergence times. Thorne *et al.*[37] and others[38–40] have proposed methods of estimating the dates associated with branch points in a phylogenetic tree. These methods can estimate when speciation or gene duplication occurred. In the past, a molecular clock assumption[41] had to be invoked when placing times or events on a tree. The molecular clock asserts that the rate of sequence evolution is identical for all species over the entire tree. The clock assumption in its strictest sense is certainly violated in most trees that include divergent taxa, but molecular evolution is usually clocklike to some extent.

In the model of Thorne *et al.*[37,38], the rate of molecular evolution evolves over the tree: species that are evolving quickly are expected to give rise to species with high, but not necessarily identical, rates of evolution. Even an intuitively simple model such as this is relatively parameter rich. In addition to the other parameters that govern the processes of molecular evolution, every branch in the tree must have a rate of evolution and time associated with it. Estimating all of these parameters simultaneously would be computationally taxing and the ML estimates would probably be unstable (because of strong interactions between similar parameters, such as the rate and time associated with a branch). Thorne *et al.*[37] have successfully implemented their model in the Bayesian context. Aris-Brosou and Yang[40] have reviewed and extended these methods of estimating divergence times, and applied them to the divergence of 18S rRNA sequences from a diverse collection of animals. Their results are more consistent with the fossil record than some molecular dating techniques, lending support to the idea that there was a Cambrian explosion that gave rise to the existing metazoan phyla.

Marginal versus joint estimation. There is a fundamental difference between the way ML and Bayesian approaches treat parameters in the models of sequence evolution. Often, these parameters are nuisance parameters; they are not of direct interest, but must be dealt with because they are found in the likelihood equations. Joint estimation — the approach most commonly used under the ML criterion — entails finding the highest point in the 'parameter landscape.' A Bayesian analysis measures the volume under a posterior-probability surface rather than its maximum height. So, the nuisance parameters are integrated out (marginalized) to obtain the marginal posterior probability of a tree. So, even if the prior-probability distribution is flat, ML and Bayesian techniques prefer different trees because of the distinction between joint and marginal estimation.

FIGURES 1a and 1b illustrate disagreements between joint and marginal estimation. For simplicity, these figures depict each tree as having its own curve through a one-dimensional parameter landscape (in reality, the geometry of tree space is of high dimensionality[42]). In both figures, joint estimation would prefer tree A (shown in blue), whereas marginalizing over the parameter x would prefer tree B (shown in orange).

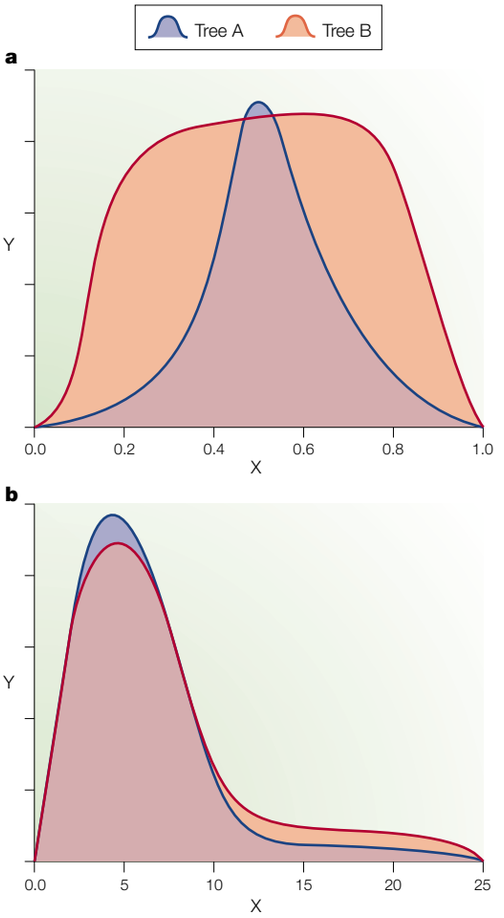


Figure 1 | **Contrast between marginal and joint estimation.** Panels **a** and **b** depict the likelihood profile for two trees versus a hypothetical parameter *x*. The *x* axis represents some nuisance parameter (for example, the ratio of the rate of transitions to the rate of transversions). The *y* axis represents the likelihood in the case of ML, or the posterior-probability density in a Bayesian approach. The area under the likelihood curve for tree A is shown in light blue, the area for tree B is shown in orange. Mauve regions are under the curve for both trees. In both cases, jointly estimating *x* and the tree favours tree A (that is, the highest peak is blue in both cases), but marginalizing over *x* favours tree B (that is, the orange area is greater than the blue area).

FIGURE 1a presents a case in which Bayesian analysis seems superior: tree A has a slightly higher peak, but tree B has good support over a wide range of values of the nuisance parameter. Marginalizing over nuisance parameters is preferable, because the estimation of parameters is imperfect; so, it is inappropriate to treat the ML estimates of parameters as the only points in parameter space that matter.

If integrating out parameters is preferable, why not marginalize over all parameters? The answer is that we integrate out parameters by weighting them according to their posterior probability, and this requires a prior probability in addition to the likelihood. Advocates of ML are uncomfortable with specifying prior distributions (which they regard as too subjective) for all parameters.

FIGURE 1b presents a concern about marginalizing over all parameters. In this example, the value of *x* seems to be <10 (this region has much higher likelihood/posterior for both trees). Tree B would be chosen by a Bayesian analysis because it has higher posterior probability for large values of *x*. It seems troubling that tree B is preferred on the basis of higher support in a clearly suboptimal region of parameter space. The Bayesian approach is not really misbehaving, even if the results seem counter-intuitive. If the data had strongly rejected values of $x > 10$, both trees would have had likelihoods near zero for that region of parameter space and there would have been no apparent problem. The example underscores the fact that every part of the surface affects the results, so careful consideration must be given to the prior distribution over the entire range of the parameter values.

When there are few parameters and a large amount of data, the debate between marginal and joint estimation is largely academic; the likelihood and posterior landscapes become steep thin spires and the height of the peak is a good predictor of the integral over the whole surface. Marginalizing becomes increasingly helpful as the amount of data decreases relative to the number of parameters (for example, when complex models are used). In these cases, the likelihood surface resembles rolling hills, and consideration of the substantial uncertainty in the values of parameters is necessary. This is also a situation in which the prior can strongly influence the analysis.

Markov chain Monte Carlo. As the previous discussion indicates, Bayesian analysis involves specifying a model and a prior distribution and then integrating the product of these quantities over all possible parameter values to determine the posterior probability for each tree. The likelihood functions for phylogenetic models are too complex to integrate analytically, so Bayesian approaches rely on MCMC[43,44] — a remarkable algorithm that is used for approximating probability distributions in a wide variety of contexts.

MCMC works by taking a series of steps that form a conceptual chain. At each step, a new location in parameter space is proposed as the next link in the chain. This proposed location is usually similar to the present one because it is generated by the random perturbation of a few of the parameters in the present state of the chain. The relative posterior-probability density at the new location is calculated. If the new location has a higher posterior-probability density than that of the present location of the chain, the move is accepted — the proposed location becomes the next link in the chain and the cycle is repeated. If the proposed location has a lower posterior-probability density, the move will be accepted only a proportion (p) of the time, where p is the ratio of the posterior of the proposed location compared with the posterior of the present location (in short, small steps downward are accepted often, whereas big leaps downward are discouraged). If the proposed location is rejected, the present location is added as the next link in the chain (so, the last two links in the chain will be identical) and the cycle is repeated. If the method for

BAYES FACTORS
The ratio of the posterior odds to the prior odds for two hypotheses of interest. Bayes factors attempt to measure how strongly the data support or refute a hypothesis.

proposing moves is not symmetrical, the rules for deciding whether to accept or reject a move must be altered[43]. By repeating this procedure millions of times, a long chain of locations in parameter space is created. The chain tends to stay in regions of high posterior probability; from these regions, almost all proposed moves are downhill and are rarely accepted. By design, the proportion of time that the chain spends in any region of parameter space can be used as an estimate of the posterior probability of that region. By creating long chains, this method of estimation can be made arbitrarily accurate.

In phylogenetics, the relevant location in parameter space is both a description of the tree and a specification of all the parameters in the model of sequence evolution. A chain is constructed that moves through different trees and models of evolution. At the end of the analysis, the researcher is given an estimate of the probability that any particular tree is the true evolutionary tree given the observed data. Of course, this probability is contingent on the model of evolution being adequate and the prior distributions on the parameters being reasonable, but it still represents an intuitive measure of the amount of confidence that should be placed in the tree.

The art of constructing a reliable MCMC sampler lies in the determination of how new locations are proposed. Almost any type of move is allowed, but some are much more effective than others. Typically, the proposed location is close to the present location of the chain. Because the posterior density is usually low for most places in parameter space, pronounced moves are likely to have low posterior probabilities. Chains that propose large moves rarely accept them and have to be run for a long time to explore the full possibilities of 'tree space'. If the proposals are too similar to the present location, the chain will accept the moves, but it will still take a long run to produce accurate results. Assessing whether or not the chain has run long enough to provide reliable estimates of the posterior probability is a crucial issue for the users of MCMC; this issue is likely to receive a great deal of research in the next few years.

Rannala and Yang[34], Larget and Simon[32], Li *et al.*[33] and Huelsenbeck and Ronquist[31] have pioneered the development of MCMC proposals for phylogenetics. Early studies indicate that it is possible to get reasonable measures of uncertainty much faster than by ML bootstrapping. Whether these results are general to most data sets (or whether approximate versions of ML bootstrapping represent an even faster alternative) has yet to be adequately addressed.

In general, Bayesian statisticians avoid the standard approach of specifying one hypothesis as the null hypothesis and asking whether or not the data are strong enough to reject it. Because the output of a Bayesian analysis is the posterior probability of any solution, standard probability rules can be used to address which hypothesis should be believed, and how strongly. For instance, if one hypothesis is consistent with 10 different trees and the alternative is consistent with all other tree shapes, the probability that the first hypothesis is correct is simply the sum of the posterior probabilities of the 10 trees. The Bayesian approach is intuitive and is particularly useful when the number of alternative hypotheses is huge (traditional hypothesis testing can be tedious in these cases). In some situations, researchers might want to reject a particular hypothesis and show conclusively that the result is not sensitive to the priors used. In this case, it is possible to either use a large range of priors or rely on BAYES FACTORS, which are less dependent on the priors that were chosen. Bayes factors give an estimate of how much the data support one hypothesis over another (see REF. 45 for interpretation guidelines).

Conclusions

The estimation of phylogenies has become a regular step in the analysis of new gene sequences. It is still too early to tell if Bayesian approaches will revolutionize tree estimation in general, but, is already clear that MCMC-based approaches are extending the field by answering previously intractable questions. These new techniques seem poised to teach us a great deal about the tree of life and molecular genetics.

1. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–502 (2000).
2. Huelsenbeck, J. P. & Bollback, J. P. Empirical and hierarchical Bayesian estimation of ancestral states. *Syst. Biol.* **50**, 351–366 (2001).
3. Metzker, M. L. *et al.* Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl Acad. Sci. USA* **99**, 14292–14297 (2002).
4. Anderson, J. F. *et al.* Isolation of West Nile virus from mosquitoes, crows, and a Cooper's hawk in Connecticut. *Science* **286**, 2331–2333 (1999).
5. Lanciotti, R. S. *et al.* Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science* **286**, 2333–2337 (1999).
6. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. in *Molecular Systematics* (eds Hillis, D. M., Moritz, C. & Mable, B. K.) 407–514 (Sinauer Associates, Sunderland, Massachusetts, 1996).
An excellent review of parsimony, ML and distance approaches to phylogenetic inference.
7. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
8. Studier, J. A. & Keppler, K. J. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* **5**, 729–731 (1988).
9. Steel, M. & Penny, D. Parsimony, likelihood and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* **16**, 839–850 (2000).
10. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, New York, 2000).
11. Takahashi, K. & Nei, M. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol. Biol. Evol.* **17**, 1251–1258 (2000).
12. Farris, J. S. Methods for computing Wagner trees. *Syst. Zool.* **19**, 83–92 (1970).
13. Fitch, W. M. Toward defining the course of evolution: minimal change for a specific tree topology. *Syst. Zool.* **20**, 406–416 (1971).
14. Kluge, A. G. & Farris, J. S. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* **18**, 1–32 (1969).
15. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410 (1978).
A seminal paper that reported the phenomenon of long-branch attraction.
16. Hillis, D. M. Inferring complex phylogenies. *Nature* **383**, 130–131 (1996).
17. Kim, J. H. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Syst. Biol.* **45**, 363–374 (1996).
18. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
19. Whelan, S., Lio, P. & Goldman, N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* **17**, 262–272 (2001).
20. Edwards, A. W. F. *Likelihood* (Oxford Univ. Press, Oxford, UK, 1972).
21. Rogers, J. S. & Swofford, D. L. A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences. *Syst. Biol.* **47**, 77–89 (1998).
22. Efron, B. Bootstrap methods: another look at the jackknife. *Annals Stat.* **7**, 1–26 (1979).
23. Felsenstein, J. Confidence intervals on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
24. Goldman, N., Anderson, J. P. & Rodrigo, A. G. Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* **49**, 652–670 (2000).
A useful taxonomy of the hypothesis-testing approaches for likelihood-based phylogenetics.

25. Hillis, D. M. & Bull, J. J. An empirical test of bootstrapping as a methods for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182–192 (1993).
26. Zharkikh, A. & Li, W.-H. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *J. Mol. Evol.* **9**, 1119–1147 (1992).
27. Felsenstein, J. & Kishino, H. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst. Biol.* **42**, 193–200 (1993).
28. Efron, B., Halloran, E. & Holmes, S. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci. USA* **93**, 13429–13434 (1996).
29. Swofford, D. L. *et al.* Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* **50**, 525–539 (2001).
A recent contribution to the debate concerning parsimony and likelihood.
30. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314 (2001).
A discussion of the promise that Bayesian phylogenetics holds for transforming evolutionary biology.
31. Huelsenbeck, J. P. & Ronquist, F. R. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
32. Larget, B. & Simon, D. L. Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**, 750–759 (1999).
33. Li, S., Pearl, D. K. & Doss, H. Phylogenetic tree construction using Markov Chain Monte Carlo. *J. Am. Stat. Assoc.* **95**, 493–508 (2000).
34. Rannala, B. & Yang, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304–311 (1996).
35. Yang, Z. H. & Rannala, B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol.* **14**, 717–724 (1997).
36. Carlin, B. P. & Louis, T. A. (eds) *Bayes and Empirical Bayes Methods for Data Analysis* (Chapman and Hall/CRC, Boca Raton, 2000).
37. Thorne, J. L., Kishino, H. & Painter, I. S. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647–1657 (1998).
38. Kishino, H., Thorne, J. L. & Bruno, W. J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**, 352–361 (2001).
39. Huelsenbeck, J. P., Larget, B. & Swofford, D. L. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**, 1879–1892 (2000).
40. Aris-Brosou, S. & Yang, Z. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst. Biol.* **51**, 703–714 (2002).
41. Zuckerkandl, E. & Pauling, L. in *Horizons in Biochemistry* (eds Kasha, M. & Pullman, B.) 189–225 (Academic Press, New York, 1962).
42. Kim, J. Geometry of phylogenetic estimation. *Mol. Phylogenet. Evol.* **17**, 58–75 (2000).
43. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
44. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
References 43 and 44 present the Metropolis–Hastings algorithm that is the underpinning of many implementations of MCMC.
45. Raftery, A. in *Markov Chain Monte Carlo in Practice* (eds Gilks, W. R., Richardson, S. & Spiegelhalter, D. J.) 163–187 (Chapman and Hall, New York, 1995).
46. Consortium, M. G. S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
47. Chang, B. S. W., Jonsson, K., Kazmi, M. A., Donoghue, M. J. & Sakmar, T. P. Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.* **19**, 1483–1489 (2002).
48. Pupko, T., Pe'er, I., Hasegawa, M., Grauer, D. & Friedman, N. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: application to the evolution of five gene families. *Bioinformatics* **18**, 1116–1123 (2002).
49. Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. Predicting the evolution of human influenza A. *Science* **286**, 1921–1925 (1999).
50. Nielsen, R. & Huelsenbeck, J. P. in *Pacific Symposium on Biocomputing* (eds Altman, R. B., Dunker, A. K., Hunter, L., Lauderdale, K. & Klein, T. E.) 576–588 (World Scientific, Singapore, 2002).
51. Anisimova, M., Bielawski, J. P. & Yang, Z. H. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* **19**, 950–958 (2002).
52. Suchard, M. A., Weiss, R. E., Dorman, K. S. & Sinsheimer, J. S. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. *Syst. Biol.* **51**, 715–728 (2002).
53. Fleming, M. A., Potter, J. D., Ramirez, C. J., Ostrander, G. K. & Ostrander, E. A. Understanding missense mutations in the *BRCA1* gene: an evolutionary approach. *Proc. Natl Acad. Sci. USA* **100**, 1151–1156 (2001).
54. Hughes, J. M., Peters, C. J., Cohen, M. L. & Mahy, B. W. Hantavirus pulmonary syndrome: an emerging infections disease. *Science* **262**, 850–851 (1993).
55. Thorne, J. L., Kishino, H. & Felsenstein, J. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* **34**, 3–16 (1992).
56. Thorne, J. L., Kishino, H. & Felsenstein, J. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**, 114–124 (1991).
57. Mitchison, G. J. A probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol.* **49**, 11–22 (1999).
58. Holmes, I. & Bruno, W. J. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17**, 803–820 (2001).
59. Lee, M. S. Y. Unalignable sequences and molecular evolution. *Trends Ecol. Evol.* **16**, 681–685 (2001).
60. Posada, D. & Crandall, K. A. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* **50**, 580–601 (2001).
61. Goldman, N. & Whelan, S. Statistical tests of γ-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **17**, 974–978 (2000).
62. Ota, R., Waddell, P. J., Hasegawa, M., Shimodaira, H. & Kishino, H. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol. Biol. Evol.* **17**, 798–803 (2000).
63. Suchard, M. A., Weiss, R. E. & Sinsheimer, J. S. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18**, 1001–1013 (2001).
64. Lewis, P. O. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.* **15**, 277–283 (1998).
65. Matsuda, H. in *Pacific Symposium on Biocomputing* (eds Hunter, L. & Klein, T. E.) 512–523 (World Scientific, London, 1996).
66. Lemmon, A. R. & Milinkovitch, M. C. The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation. *Proc. Natl Acad. Sci. USA* **99**, 10516–10521 (2002).
67. Salter, L. A. & Pearl, D. K. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Syst. Biol.* **50**, 7–17 (2001).
68. Nixon, K. C. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* **15**, 407–414 (1999).

Acknowledgements

This manuscript was greatly improved by comments from three anonymous reviewers. The authors gratefully acknowledge the financial support provided by a grant from the Alfred P. Sloan Foundation/National Science Foundation awarded to P.O.L.

Online links

DATABASES

The following term in this article is linked online to:
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
BRCA1

FURTHER INFORMATION

BAMBE: http://www.mathcs.duq.edu/larget/bambe.html
BLAST: http://www.ncbi.nlm.nih.gov/BLAST
MEGA: http://www.megasoftware.net
MrBayes: http://morphbank.ebc.uu.se/mrbayes
PAML: http://abacus.gene.ucl.ac.uk/software/paml.html
PAUP: http://paup.csit.fsu.edu/index.html
PHYLIP: http://evolution.genetics.washington.edu/phylip.html
Phylogeny programs:
http://evolution.genetics.washington.edu/phylip/software.html
Access to this interactive links box is free online.