Maximum Likelihood Methods for Phylogenetic Inference

Amrit Dhar[1] and Vladimir N. Minin[1,2]

[1]Department of Statistics, University of Washington, Seattle
[2]Department of Biology, University of Washington, Seattle

December 29, 2015

Abstract

In this article, we provide an overview of maximum likelihood methods for phylogenetic inference. A brief introduction to general maximum likelihood estimation is provided. We define a phylogenetic likelihood, summarize how to compute this likelihood, and then discuss approaches used to maximize the phylogenetic likelihood function. We discuss a property of the maximum likelihood estimation, called consistency, that states that the maximum likelihood phylogeny will converge to the true phylogenetic tree with as more and more data are added to the analysis. We describe the bootstrap, a popular technique used to characterize the uncertainty in parameter estimates, and then outline its use in phylogenetic maximum likelihood estimation. A short example is given to illustrate the use of phylogenetic maximum likelihood techniques on a real dataset of primate mitochondrial DNA sequences.

Keywords: evolutionary trees; maximum likelihood estimation; optimization; consistency; bootstrap

1 Introduction

Maximum likelihood estimation is an extremely popular statistical inference framework that is used to estimate the parameters in a probabilistic data generating model. This conceptually simple method provides parameter estimates that have good statistical properties. Before delving into maximum likelihood techniques for phylogenetic tree reconstruction, we present a simple example of maximum likelihood estimation to illustrate how this procedure works.

Suppose we have a coin that has some unknown probability p of landing on heads. We are interested in estimating this unknown probability based on the outcomes of observed coin flips. For example, one could imagine flipping this coin independently 10 times, resulting in an observed sequence of $HTTHTHTHTT$, denoted by D. The maximum likelihood principle suggests that the best guess for p is obtained by choosing p so that the probability of observing the sequence $HTTHTHTHTT$ is the highest. The probability of observing the sequence $HTTHTHTHTT$ from this coin is:

$$L(p) = \Pr(D; p) = p(1-p)(1-p)p(1-p)p(1-p)p(1-p)(1-p) = p^4(1-p)^6. \tag{1}$$

We are able to take a product of p and $1-p$ terms because of the independence of the coin flips. Equation (1) is referred to as the likelihood given parameter p . Because we are picking the value of p that makes $L(p)$ the largest, this likelihood function is treated as a function of p, not as a function of the data D. From basic calculus, we know that finding the p that maximizes a differentiable function w.r.t. p is equivalent to solving the equation $L'(p) = 0$, which is:

$$2(-1+p)^5 p^3(-2+5p) = 0. \tag{2}$$

This results in three roots $p = 0, 0.4, 1$. A quick inspection of the likelihood function, depicted in Figure 1, evaluated at all three roots shows that $p = 0.4$ is the maximum likelihood estimate of p. This is an intuitive estimate because it is just the proportion of heads observed in the sequence $HTTHTHTHTT$. It turns out that this maximum likelihood estimator (MLE) of p tends to the true, unknown value of p as we observe more and more coin flips — this is called consistency in statistics. Also, this estimator uses available data in the most optimal way, again as we observe more and more coin flips – this is called efficiency in statistics. Under fairly general and reasonable regularity conditions, consistency and efficiency hold for a large class of maximum likelihood estimators (van der Vaart, 1998). This easy-to-understand estimation principle along with the associated optimality properties for a wide class of likelihood models make maximum likelihood an attractive procedure for many parameter estimation problems, including the problem of estimating phylogenetic relationships from molecular sequence data.

2 Phylogenetic Likelihood

2.1 Likelihood Description

Throughout this presentation, we restrict our attention to DNA sequence data for simplicity (although the methods we describe are compatible with other discrete character datasets as well). Suppose we observe m aligned sequences of DNA (potentially corresponding to m distinct species), where each sequence has nucleotide observations recorded at n distinct sites. Gaps in the alignment are usually treated as missing data, but more accurate treatment of insertions and deletions is possible (Redelings and Suchard, 2005; Liu et al., 2012). Just as we considered the likelihood of the observed coin flips as a function of the unknown parameter p, here we will examine the likelihood of the DNA sequence data as a function of the unknown tree topology and branch lengths. Given a tree topology with branch lengths, we can use a substitution model to calculate the probabilities of state changes along the branches of the tree; substitution models are described in great detail in the "Models and Model Selection" article contained within this encyclopedia. Specifically if t denotes a branch length on a tree, substitution models allow us to calculate $p_{ij}(t)$, which denotes the probability of going from state i to state j on a branch of length t, where $i, j \in \{A, G, C, T\}$. Note that branch lengths are commonly measured in expected number of substitutions per site, not in clock time, because estimating substitution rates and branch lengths in units of clock time requires additional information about branching and/or sampling times in the phylogeny (Drummond et al., 2006).

Two assumptions are made that are crucial to the rest of the analysis (Felsenstein, 2004):

1. Evolution at different sites (on a given tree) is independent.

2. Conditional on the internal node states, evolution proceeds independently on different branches of the phylogeny.

Let $L(\tau, \mathbf{t}, \boldsymbol{\theta})$ be the likelihood corresponding to the $m \times n$ DNA sequence alignment matrix $\mathbf{y}$ for a given tree topology τ with branch length vector $\mathbf{t}$ and substitution model parameter vector $\boldsymbol{\theta}$. We can write the likelihood as:

$$L(\tau, \mathbf{t}, \boldsymbol{\theta}) = \Pr(\mathbf{y}; \tau, \mathbf{t}, \boldsymbol{\theta}) = \prod_{i=1}^{n} P(\mathbf{y}_i; \tau, \mathbf{t}, \boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\tau, \mathbf{t}, \boldsymbol{\theta}), \tag{3}$$

where $\mathbf{y}_i$ is the $m \times 1$ vector of observed nucleotides at the i'th site and $L_i(\tau, \mathbf{t}, \boldsymbol{\theta})$ is the site i likelihood. This factorization follows directly from the first independence assumption given above. Thus, we can find the likelihood of the whole sequence matrix by finding the likelihoods for each of the n sites. Suppose we observed the nucleotide vector (A, T, C, T) at a particular site (assuming there were only $m = 4$ aligned sequences). We will use the example tree τ_{ex} given in Figure 2 to help illustrate how to calculate the likelihood at this site.

Using the tree τ_{ex}, we can deconstruct the likelihood of this nucleotide vector in the following manner:

$$\begin{aligned} \Pr(A, T, C, T; \tau_{ex}, \mathbf{t}_{ex}, \boldsymbol{\theta}) &= \sum_{x} \sum_{y} \sum_{z} \Pr(A, T, C, T, x, y, z; \tau_{ex}, \mathbf{t}_{ex}, \boldsymbol{\theta}) \\ &= \sum_{x} \sum_{y} \sum_{z} \pi_x p_{xy}(t_1) p_{xz}(t_2) p_{yA}(t_3) p_{yT}(t_4) p_{zC}(t_5) p_{zT}(t_6), \end{aligned} \tag{4}$$

where the summations are over the elements in $\{A, G, C, T\}$. We obtain equation (4) by conditioning on the internal node states and by invoking the assumption of independent evolution across branches. Note that, following common practice, we assumed that the initial distribution at the root of the phylogeny is $\boldsymbol{\pi} = (\pi_A, \pi_G, \pi_C, \pi_T)^T$, the stationary distribution of the substitution model (Page and Holmes, 2009). Looking at equation (4), it is clear that we will need to take a sum over $4^3 = 64$ probabilities. For only $m = 4$ terminal nodes, this computation is reasonable; as m grows, the computation becomes problematic because the sum will involve 4^{m-1} terms over $m - 1$ internal nodes. In the next section, we describe an algorithm that can efficiently calculate this site likelihood by eliminating redundant computations.

2.2 Likelihood Computation

First presented by Felsenstein (1973), the pruning algorithm is a standard technique used to efficiently compute phylogenetic likelihoods. This algorithm is a particular type of dynamic programming technique and takes advantage of the distributive law of algebra to achieve efficiency gains. For example, the expression $ab + ac + ad + ae$, where a, b, c, d, e are scalars, requires 7 computations (4 multiplications and 3 additions). By noting the scalar a appears in all 4 multiplications, we can use the distributive law to re-write the above expression as $a(b + c + d + e)$, which requires 4 computations (3 additions and 1 multiplication). Thus by applying the distributive law to expressions that contain sums of product terms, like in equation (4), we can reduce the number of computations required to evaluate them.

For the phylogenetic site likelihood given in equation (4), we can utilize the distributive law by pushing the summations as far right as possible:

$$\begin{aligned} P(A,T,C,T|\tau_{ex},\mathbf{t}_{ex},\boldsymbol{\theta}) &= \sum_x \sum_y \sum_z \pi_x p_{xy}(t_1)p_{xz}(t_2)p_{yA}(t_3)p_{yT}(t_4)p_{zC}(t_5)p_{zT}(t_6) \\ &= \sum_x \pi_x \left[\sum_y p_{xy}(t_1)p_{yA}(t_3)p_{yT}(t_4)\right]\left[\sum_z p_{xz}(t_2)p_{zC}(t_5)p_{zT}(t_6)\right]. \end{aligned} \tag{5}$$

This formulation suggests calculating $\sum_y p_{xy}(t_1)p_{yA}(t_3)p_{yT}(t_4)$ and $\sum_z p_{xz}(t_2)p_{zC}(t_5)p_{zT}(t_6)$ first, caching these intermediate results for all possible values x, and then computing the final sum over x. Note that we can visualize this procedure as traversing τ_{ex} bottom-up because the sums over y and z are evaluated before the sum over x. Instead of naively summing over $4^3 = 64$ terms, this reformulation requires a sum over $4 \times 3 = 12$ terms. For an arbitrary number of terminal nodes m, a similar reformulation will reduce the number of computations from being exponential in m to being linear in m.

The nested, bottom-up nature of the above computation leads naturally to a recursion for calculating site likelihoods. We define $L(k,i)$ to be the conditional likelihood of a subtree with root node k being in state i. Conceptually, it is the likelihood of the observed terminal nodes below node k, conditional on node k being in state i. For example in τ_{ex}, $L(y,i)$ represents the likelihood of observing (A,T) below node y, conditional on node y being in state i. For any internal node k (in state i) with children v,w and corresponding branch lengths t_v,t_w, Felsenstein (1973) defines the recursion to be:

$$L(k,i) = \left[\sum_{s_1\in\{A,G,C,T\}} p_{is_1}(t_v)L(v,s_1)\right]\left[\sum_{s_2\in\{A,G,C,T\}} p_{is_2}(t_w)L(w,s_2)\right] \tag{6}$$

for all states $i \in \{A,G,C,T\}$. This recursion uses the conditional likelihoods calculated for the children of node k to compute the conditional likelihoods for node k itself.

The derivation of the above recursion is based on the assumption of independent evolution across different lineages and the law of total probability. The former justifies decomposition of $L(k,i)$ into a product of the two terms, enclosed in square brackets, in Equation (6), where each term represents a conditional likelihood component from one of the two lineages below node k. The component from child v is computed by summing over all possible states ($s_1 \in \{A,G,C,T\}$) to which state i could have changed to and for each possible state computes the probability of changing to that state (i.e. $p_{is_1}(t_v)$) times the probability of everything that is observed below node v, given that node v's state is s_1. A similar reasoning can be used to describe the component corresponding the other child of k — node w.

To turn the recursion in Equation (6) into an algorithm, we need to initialize the recursion by defining the $L(k,i)$ values for all terminal nodes. For any given terminal node k, $L(k,i)$ is set to 1 when state i is the observed state and 0 otherwise. This initialization process can be adjusted to account for situations when data are either missing or partially observed at the terminal nodes (Felsenstein, 2004). Once all of the terminal node $L(k,i)$'s are initialized, we continue calculating conditional likelihoods for internal nodes up the tree, computing them only if the corresponding conditional likelihoods for their children have been computed. The procedure ends after calculating the conditional likelihoods at the root ($L(\text{root},i)$, for all i); the resulting site likelihood is computed as $\sum_i \pi_i L(\text{root},i)$, where $i \in \{A,G,C,T\}$.

2.3 Root Invariance and the Pulley Principle

Before discussing the specifics of maximum likelihood phylogeny estimation, it is essential to understand the types of phylogenetic trees that will be estimated. Even though it may seem like we are estimating rooted phylogenies, it turns out that under our assumption of substitution model reversibility and without further assumptions or external information maximum likelihood methods can only estimate unrooted trees. This result is a direct consequence of the Pulley Principle, first discussed in (Felsenstein, 1981). Under the assumptions of a reversible substitution model, unconstrained branch lengths, and a root nucleotide distribution at equilibrium, the Pulley Principle states that the root may be placed anywhere on the tree without affecting the likelihood. This implies that the root is unidentifiable using likelihood methods for phylogeny estimation because the likelihood is invariant to the placement of the root. In fact, we are not estimating a single rooted tree, but an equivalence class of rooted trees that corresponds to a unique unrooted tree (Felsenstein, 1981). In Figure 3, we present two rooted trees that lie in the same equivalence class; the unrooted tree that corresponds to this equivalence class is shown below the two rooted trees. As we shift the root node associated with x, the tree topology and branch length parameters change, but the likelihood value remains the same.

3 Likelihood Maximization

3.1 Branch Length Optimization

Now that we have described how phylogenetic likelihoods can be computed, we will begin discussing how to maximize these functions with respect to the tree topology τ, branch lengths $\mathbf{t}$, and substitution model parameters $\boldsymbol{\theta}$. Phylogenetic maximum likelihood algorithms proceed by iterating between two major algorithmic steps: 1) for a given tree topology, find optimal branch lengths (i.e. the branch lengths that make the observed data most likely) and substitution model parameters 2) obtain a tree topology that maximizes the likelihood given branch lengths and substitution model parameters. We start with the continuous optimization problem.

The problem of optimizing the phylogenetic likelihood function, or equivalently the log-likelihood, over branch lengths and substitution model parameters falls into a class of nonlinear, non-convex optimization problems. This means that no existing optimization algorithm can guarantee to solve this problem. However, in practice, such problem can be solved by a myriad of hill climbing optimization methods, such as Newton-Raphson method and the expectation-maximization (EM) algorithms. Although computational ingredients for the Newton-Raphson (Schadt et al., 1998; Kenney and Gu, 2012) and the EM algorithm (Holmes and Rubin, 2002; Hobolth and Jensen, 2005) are available, often simpler methods are preferred. For example, many implementations of the phylogenetic maximum likelihood estimation update branch lengths one at a time, rather than jointly (Guindon and Gascuel, 2003).

In addition to lack of optimization convergence guarantees, there is no theory that says that a phylogenetic likelihood will have a unique maximum, although multiple maxima are rarely seen in practice (Felsenstein, 2004). Steel (1994) provides an example that shows maximum likelihood branch lengths for a given tree are not necessarily unique. Chor et al. (2000) found sequence alignments that have multiple branch length optima on the same tree. In contrast, Rogers and Swofford (1999) performed numerous simulation studies and concludes that it's extremely unlikely that maximum likelihood estimation results in multiple local optima for a given phylogenetic tree. Given these intriguing results, it is not surprising that studying properties of the phylogenetic likelihood function remains an active research area.

3.2 Tree Topology Search

Because there exist finitely many tree topologies, we could, in principle, optimize branch lengths and substitution parameters for every possible tree topology and choose the tree that had the highest likelihood value as the maximum likelihood tree. Unfortunately, the set of possible topologies is extremely large (Felsenstein, 1981) so naively searching over this tree space is computationally infeasible. Various heuristics are used to find the topology that has the highest likelihood. All these methods use local modifications of the previously visited tree topologies to find a new tree with a higher likelihood. For example, early methods, such as PHYLIP (Felsenstein, 1989) and PAUP* (Swofford, 2003) packages, traverse the tree topology space greedily by comparing the likelihood values between these modified trees and by choosing the topology that increases the likelihood the most; the procedure will end if there are not any trees that increase the likelihood. While this local tree search is faster than the exhaustive search over all possible trees, it is still inefficient because it has to optimize branch lengths and evaluate likelihoods for all of the rejected trees (Guindon and Gascuel, 2003).

Many other tree search heuristics have been introduced to improve upon the aforementioned hill-climbing methods. Salter and Pearl (2001) proposed a stochastic search algorithm that uses simulated annealing to move through tree space. This stochastic search was found to be faster and less likely to become trapped in local optima, when compared to PHYLIP and PAUP, for several simulated and real data examples. The improvement in speed is largely due to the fact that stochastic search algorithms gradually optimize branch lengths and other model parameters as the tree search goes on and avoid the full optimization steps used within hill-climbing methods for every candidate tree (Salter and Pearl, 2001). In addition, Guindon and Gascuel (2003) presented a simple hill-climbing algorithm that adjusts the tree topology and branch lengths simultaneously. By performing joint optimization, this procedure tends not to get stuck at local modes of the likelihood and produces extremely accurate estimates of the tree topology (Guindon and Gascuel, 2003). Although these optimization strategies work well in many practical situations, it is important to keep in mind that all local hill climbing optimization methods are prone to getting stuck in local maxima and not returning the true MLE phylogeny as a result.

4 Consistency of Maximum Likelihood Estimates

An important question of interest is whether the maximum likelihood phylogenetic estimation process is able to reconstruct the true phylogeny as we gather more and more data. More precisely, as we collect data at more and more sites for a fixed set of sequences, will the maximum likelihood phylogeny estimates tend to the true phylogenetic tree? As we briefly mentioned before, consistency holds for a wide class of maximum likelihood estimators under sufficiently broad regularity conditions (van der Vaart, 1998). Despite this, consistency for maximum likelihood phylogenies has been hard to establish due to the complex nature of the parameter space.

Felsenstein (1973) suggested that consistency could be proven for maximum likelihood phylogeny estimates by using a modified version of a general consistency proof found in (Wald, 1949), although the proof was never explicitly given. Chang (1996) presented one of the earliest proofs of consistency, but did not consider the branch length parameters in his setup. RoyChoudhury et al. (2015) provide a complete proof of consistency by verifying the Wald conditions in this setting; their proof is dependent on a previously constructed metric space for phylogenetic trees (Billera et al., 2001). Thus, RoyChoudhury et al. (2015) show that consistency will hold for maximum likelihood phylogenies assuming the correct model of evolution is used. A discussion regarding phylogenetic MLE consistency under model misspecification can be found in (Felsenstein, 2004).

5 Bootstrap for Phylogenies

In the process of phylogenetic estimation, it is natural to wish to quantify uncertainty about the reconstructed phylogeny. In this section, we explain how nonparametric bootstrap, a general statistical technique for assessing sampling variability of an estimator (Efron, 1979), can be used to assign confidence to phylogenetic MLEs. Before getting into the details of the bootstrap procedure, it is instructive to imagine what would be a statistically ideal way to estimate phylogenetic uncertainty. Suppose we could repeatedly "re-run" evolution along the same true phylogenetic history to obtain replicated molecular sequence alignments. Then, estimating phylogenies based on these alignments via maximum likelihood would tell us something about sampling variability of our estimation. For example, if the phylogenies obtained from the replicated re-runs of evolution were nearly identical, we would conclude that our phylogenetic estimation is very precise, enjoying low sampling variability. The idea behind bootstrap is to get around re-running evolution, which is clearly infeasible, by resampling the observed data.

5.1 Bootstrap Description

The bootstrap procedure starts by generating B replicate datasets. Each dataset is obtained by repeatedly sampling n alignment sites with replacement (i.e. sampling columns). Maximum likelihood estimation is then applied to each of these B bootstrapped sequence alignments, and uncertainty is assessed by summarizing the similarities between the bootstrapped phylogenies (Felsenstein, 1985). Figure 4 displays some bootstrap datasets drawn from an observed sequence alignment; note that each dataset consists of randomly sampled columns from this observed alignment. While bootstrap sampling is conceptually simple, summarizing bootstrapped trees is not an easy task so we devote the next subsection to this topic. For a more detailed account of bootstrap sampling, please see the "Measures of Tree Support" article contained within this encyclopedia.

5.2 Summarizing Bootstrapped Phylogenies

We are interested in understanding how to summarize the similarities among the B bootstrap trees because a high degree of similarity corresponds to a low degree of variability. For instance, if a particular tree split — a bipartition of the species set — appears in most of the bootstrap trees, then we will be more confident about that tree split being in the true phylogeny. One way to keep track of which tree splits appear most often in the bootstrap trees is to construct a majority-rule consensus tree — a tree that is constructed from tree splits that appear in a majority of the bootstrap trees. This is done by first enumerating all of the tree splits that occur on the bootstrap trees and then retaining only those splits that appear in more than 50% of the trees. We can then construct the consensus tree using this remaining collection of splits. In building the consensus tree, the procedure always avoids putting two splits that might conflict on the same tree because there will always be at least one bootstrap tree where the two splits coexist (Felsenstein, 1985). Thus, it is guaranteed that a consensus tree can be built from these tree splits. Once the tree is constructed, it is common to label each split with the percentage of bootstrap trees it appears in. This helps us understand which parts of the consensus tree we should have strong or weak confidence in. Figure 5 illustrates how this tree building process works by constructing a majority-rule consensus tree for a simple collection of trees. For a more comprehensive treatment of consensus trees, please see the "Consensus Trees" article contained within this encyclopedia.

6 Maximum Likelihood and more complex models of evolution

So far, we have limited our discussion of phylogenetic maximum likelihood estimation to the very basic models of molecular evolution. However, assuming more complex models does not significantly change the maximum likelihood machinery. For example, it is widely recognized that when modeling evolution of molecular sequences it is important to account for a possibility of different substitution rates across sites (Yang, 1994). Therefore, most implementations of phylogenetic maximum likelihood estimation include models that deal with such heterogeneity. Model extensions that relax the assumption of site independence (Hobolth, 2008) and that impose constraints on substitution rates across branches of the phylogeny (Rambaut and Bromham, 1998) are also possible. Increasing model complexity may eventually lead to a situation where maximum likelihood estimation ceases to produce a unique solution, so studying identifiability of models becomes an important avenue of research (Rhodes and Sullivant, 2012).

7 Example: Primate Phylogeny Estimation

In this section, we present an example of maximum likelihood phylogeny estimation applied to an alignment of mitochondrial DNA sequeces from 7 different primate species: human, chimpanzee, bonobo, gorilla, Bornean orangutan, Sumatran orangutan, and gibbon (Yang et al., 1998). The length of the alignment is 9,993 sites. We are interested in understanding the ancestral relationships among these species. We use a general time-reversible (GTR) substitution model, with a gamma distribution used to model rate variation across sites. After finding the maximum likelihood phylogeny estimate using the PhyML package (Guindon et al., 2010), we perform the bootstrap and obtain 1000 bootstrapped trees. We carry out this procedure on the full sequence alignment and on an alignment that we constructed by randomly subsampling 500 sites from the original alignment.

Figure 6 displays the maximum likelihood trees with corresponding labeled bootstrap percentages; note that the majority-rule consensus trees have been omitted from Figure 6 as they coincide with the maximum likelihood trees in this example. However, maximum likelihood trees and bootstrap consensus trees could be different, which usually happens when maximum likelihood phylogenetic inference is not very precise. In this example, maximum likelihood trees of both the full and subsampled alignment share the same topology and have similarly sized bootstrap percentages. However, notice that artificially reducing the size of the data yields more uncertainty reflected in lower bootstrap support of internal branches of the reconstructed phylogeny on the righthand side of Figure 6. Our analysis suggests an accepted relationship among the primates, with bonobos and chimpanzees being most closely related to humans (Mailund et al., 2014).

References

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, 27(4):733–767.

Chang, J. T. (1996). Full Reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency. *Mathematical Biosciences*, 137(1):51–73.

Chor, B., Hendy, M. D., Holland, B. R., and Penny, D. (2000). Multiple Maxima of Likelihood in Phylogenetic Trees: An Analytic Approach. *Molecular Biology and Evolution*, 17(10):1529–1541.

Drummond, A. J., Ho, S. Y., Phillips, M. J., Rambaut, A., et al. (2006). Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, 4(5):699.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Zoology*, 22(3):240–249.

Felsenstein, J. (1981). Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution*, 17(6):368–376.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, pages 783–791.

Felsenstein, J. (1989). PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

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. *Systematic Biology*, 59(3):307–321.

Guindon, S. and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5):696–704.

Hobolth, A. (2008). A Markov Chain Monte Carlo Expectation Maximization Algorithm for Statistical Analysis of DNA Sequence Evolution with Neighbor-Dependent Substitution Rates. *Journal of Computational and Graphical Statistics*, 17(1):138–162.

Hobolth, A. and Jensen, J. L. (2005). Statistical Inference in Evolutionary Models of DNA Sequences via the EM Algorithm. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1–22.

Holmes, I. and Rubin, G. (2002). An Expectation Maximization Algorithm for Training Hidden Substitution Models. *Journal of Molecular Biology*, 317(5):753–764.

Kenney, T. and Gu, H. (2012). Hessian Calculation for Phylogenetic Likelihood based on the Pruning Algorithm and its Applications. *Statistical Applications in Genetics and Molecular Biology*, 11(4):1–46.

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., and Linder, C. R. (2012). SATÈ-II: Very Fast and Accurate Simultaneous Estimation of Multiple Sequence Alignments and Phylogenetic Trees. *Systematic Biology*, 61(1):90–106.

Mailund, T., Munch, K., and Schierup, M. H. (2014). Lineage Sorting in Apes. *Annual Review of Genetics*, 48:519–535.

Page, R. D. and Holmes, E. C. (2009). *Molecular Evolution: A Phylogenetic Approach*. John Wiley & Sons.

Rambaut, A. and Bromham, L. (1998). Estimating Divergence Dates from Molecular Sequences. *Molecular Biology and Evolution*, 15(4):442–448.

Redelings, B. D. and Suchard, M. A. (2005). Joint Bayesian Estimation of Alignment and Phylogeny. *Systematic Biology*, 54(3):401–418.

Rhodes, J. A. and Sullivant, S. (2012). Identifiability of Large Phylogenetic Mixture Models. *Bulletin of Mathematical Biology*, 74(1):212–231.

Rogers, J. S. and Swofford, D. L. (1999). Multiple Local Maxima for Likelihoods of Phylogenetic Trees: A Simulation Study. *Molecular Biology and Evolution*, 16(8):1079–1085.

RoyChoudhury, A., Willis, A., and Bunge, J. (2015). Consistency of a Phylogenetic Tree Maximum Likelihood Estimator. *Journal of Statistical Planning and Inference*, 161:73–80.

Salter, L. A. and Pearl, D. K. (2001). Stochastic Search Strategy for Estimation of Maximum Likelihood Phylogenetic Trees. *Systematic Biology*, 50(1):7–17.

Schadt, E. E., Sinsheimer, J. S., and Lange, K. (1998). Computational Advances in Maximum Likelihood Methods for Molecular Phylogeny. *Genome Research*, 8(3):222–233.

Steel, M. (1994). The Maximum Likelihood Point for a Phylogenetic Tree Is Not Unique. *Systematic Biology*, 43(4):560–564.

Swofford, D. L. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

Yang, Z. (1994). Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *Journal of Molecular Evolution*, 39(3):306–314.

Yang, Z., Nielsen, R., and Hasegawa, M. (1998). Models of Amino Acid Substitution and Applications to Mitochondrial Protein Evolution. *Molecular Biology and Evolution*, 15(12):1600–1611.

Figures

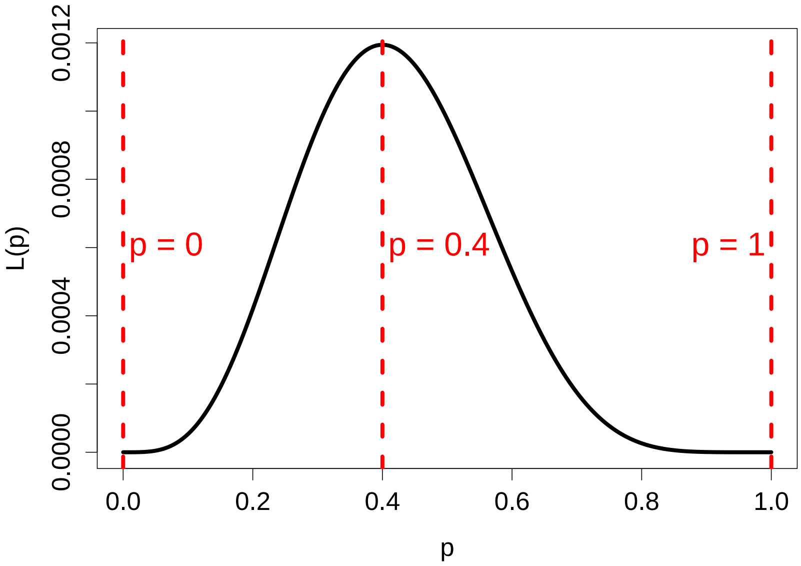


Figure 1: The likelihood of observing 4 heads in 10 coin tosses as a function of the unknown probability of heads p in a single toss. The likelihood function $p^4(1-p)^6$ is maximized at $p = 0.4$, which corresponds to the observed proportion of heads. The red dashed lines represent critical points found by differentiating the likelihood function and setting it equal to 0.

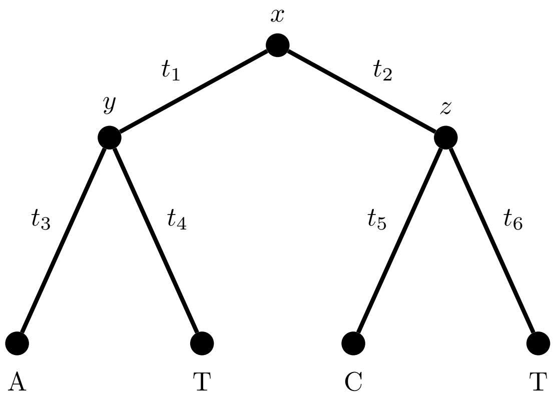
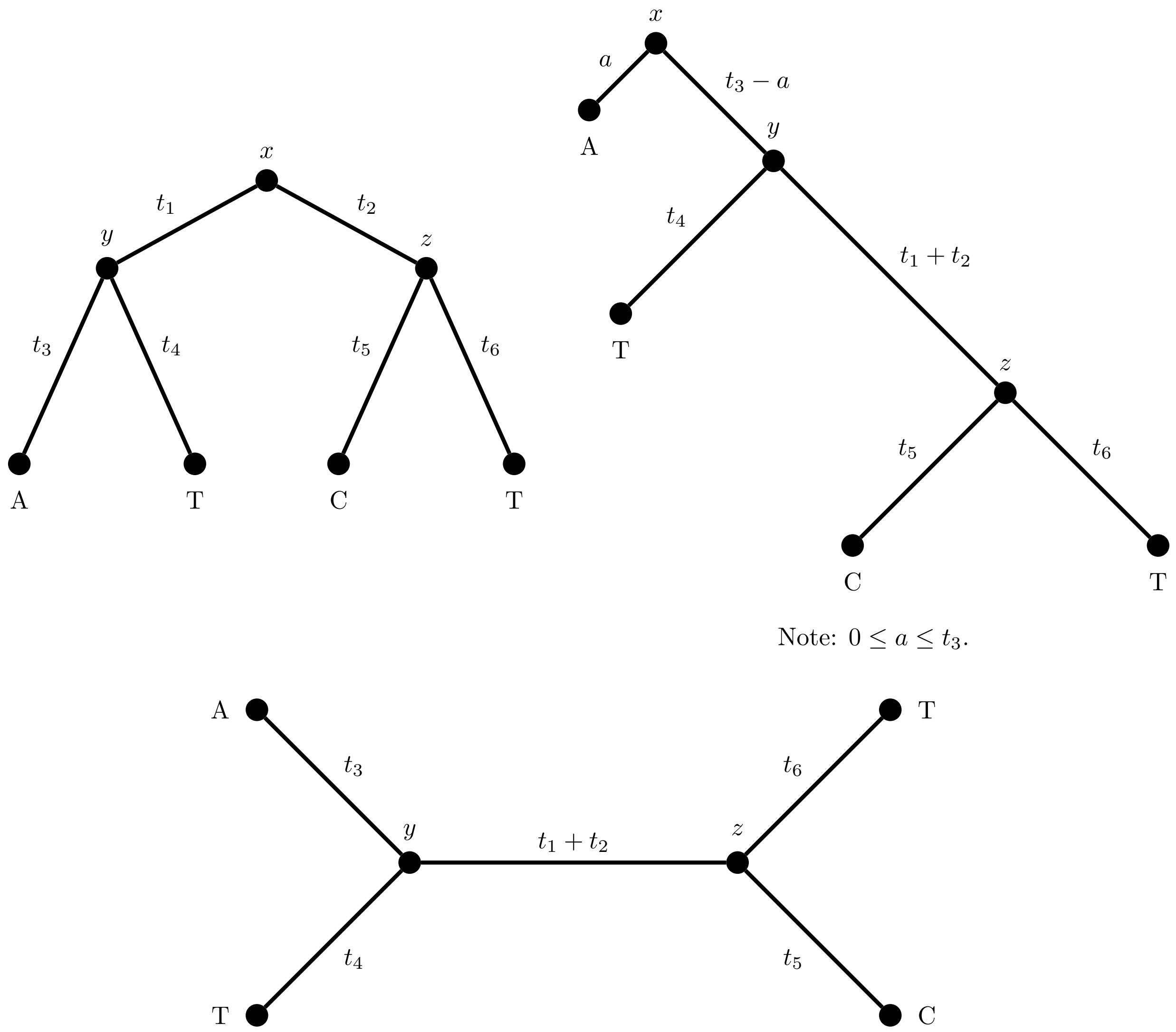


Figure 2: An example phylogenetic tree. Letters x, y, z represent the unobserved internal node states where x is associated with the root node, τ_{ex} specifies the tree topology, and $\mathbf{t}_{ex} = (t_1, t_2, ..., t_6)$ denotes the vector of branch lengths. Given the tree topology τ_{ex} and branch lengths $\mathbf{t}_{ex}$, we can calculate the likelihood of observing the nucleotide vector (A, T, C, T).



Note: $0 \leq a \leq t_3$.

Figure 3: Three trees with equivalent likelihood values under the assumptions of the Pulley Principle. The Pulley Principle states that the root of a tree may be placed anywhere on the tree without affecting the likelihood value. The two rooted trees (top) are contained within a larger equivalence class of rooted trees that uniquely corresponds to the given unrooted tree (bottom).

A	A	G	T	C	A	T	C	T	C
G	C	T	A	A	G	G	T	C	A
T	C	A	T	T	T	G	A	G	T
T	A	G	C	T	C	A	G	G	G

Observed sequence alignment ($m = 4, n = 10$)

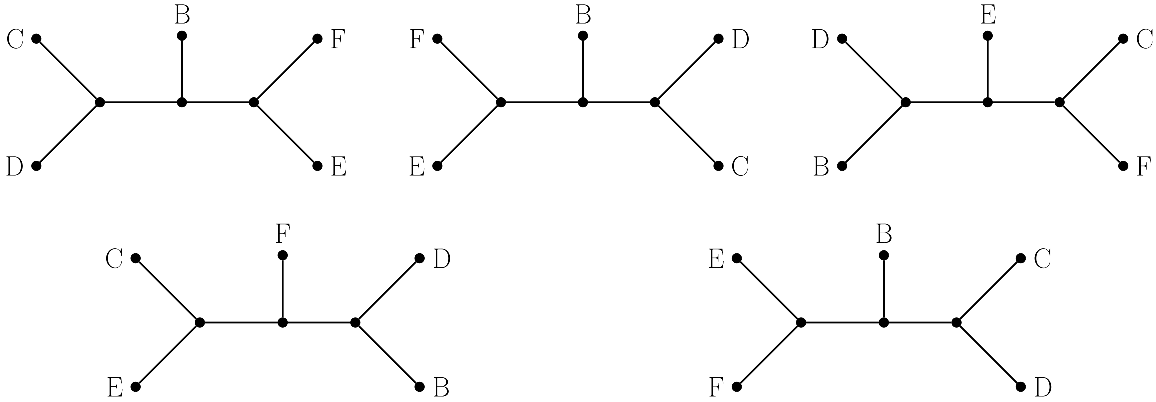
C	T	A	C	T	T	A	T	G	C
A	C	G	A	G	A	C	G	T	T
T	G	T	T	G	T	C	G	A	A
T	G	T	T	A	C	A	A	G	G

Bootstrap sample #1

C	A	A	A	T	A	A	C	C	T
T	G	C	G	C	G	C	T	A	A
A	T	C	T	G	T	C	A	T	T
G	C	A	T	G	T	A	G	G	C

Bootstrap sample #2

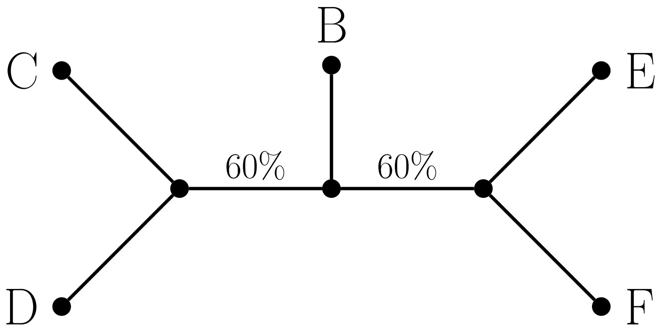
Figure 4: An example of bootstrap sampling for sequence alignment data. We obtain bootstrap datasets by randomly sampling $n = 10$ sites (i.e. columns) with replacement from the observed sequence alignment (top). We provide two examples of possible bootstrap datasets (bottom).



Tree collection

Tree Splits	Counts
$\{C, D\}\|\{B, E, F\}$	3
$\{B, C, D\}\|\{E, F\}$	3
$\{B, D\}\|\{C, E, F\}$	2
$\{B, D, E\}\|\{C, F\}$	1
$\{C, E\}\|\{B, D, F\}$	1

Number of appearances of each tree split



Majority-rule consensus tree

Figure 5: Consensus tree example with 5 species $\{B, C, D, E, F\}$. At the top of the figure, we display a sample collection of trees from which we'll build a consensus tree. A table enumerating all possible tree splits from this collection and their respective counts is given in the middle of the figure. At the bottom of the figure, we display the only consensus tree that can be constructed from the majority tree splits, which are $\{C, D\}|\{B, E, F\}$ and $\{B, C, D\}|\{E, F\}$. Note that we label each tree split with the corresponding percentage of the tree collection it appears in.

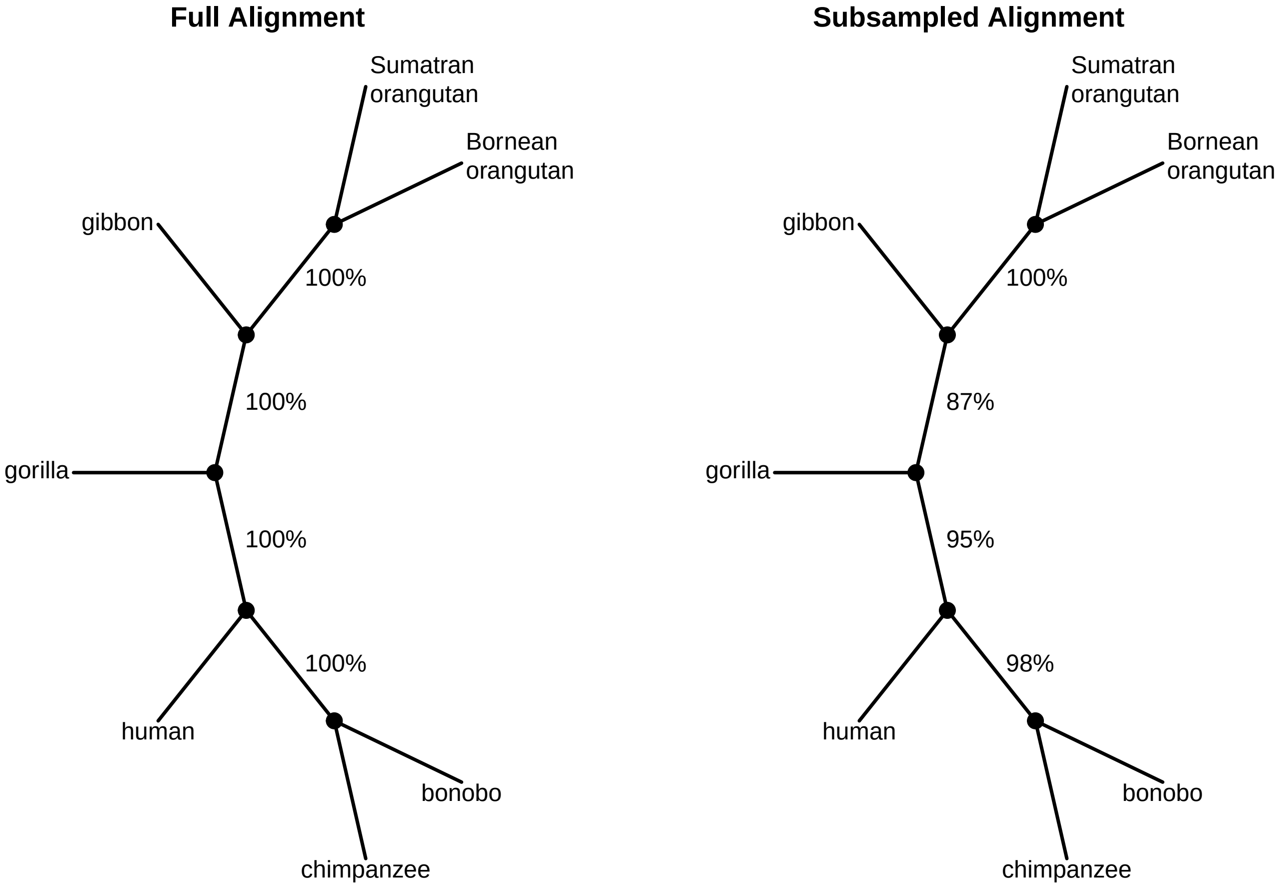


Figure 6: Maximum likelihood phylogenies with corresponding bootstrap percentages for the primate example using both the full sequence alignment and a randomly subsampled alignment containing 500 sites.

Glossary

Dynamic programming: A general method for solving complex problems by breaking them down into a collection of simpler subproblems that are similar in structure to the original problem.

Hessian matrix: A square matrix of second order partial derivatives of a real-valued function. It is often needed within numerical optimization techniques.

Hill climbing: A generic optimization strategy that relies on local searches when trying to find the global maximum of a given function. Hill climbing procedures incrementally modify variables in an optimization problem and check to see if the modified variables have achieved a higher function value (i.e. "climbed the hill").

Law of total probability: A probability formula that states how to calculate the probability of an event given a partition of the sample space.

Metric space: A set of objects (i.e. real numbers, phylogenies, etc.) equipped with a distance between any two elements.

Reversible substitution model: A Markov substitution model that, if started in equilibrium distribution, can be run backwards in time, with the resulting backward Markov model following the same probability law as the original forward model.

Simulated annealing: A probabilistic method for approximating the global maximum of a given function. It is a useful technique for avoiding getting stuck in local maxima.

Stationary distribution: A marginal probability distribution over the states of a substitution model that can be interpreted as a long-run steady state distribution as evolution occurs over time.

Substitution model: A model that specifies probabilities of state changes for DNA (or amino acid) sequence data along the branches of a given phylogeny.

Relevant Software

PhyML: `http://www.atgc-montpellier.fr/phyml/`. This software tool was used to estimate the phylogenies in our primate example. This program works with both nucleotide and protein sequence data, has an easy-to-use online interface, and can handle a wide variety of substitution models, rate heterogeneity across sites, and the phylogenetic bootstrap.

RAxML: `http://sco.h-its.org/exelixis/web/software/raxml/`. This program uses parallel processing and a simulated annealing algorithm to find maximum likelihood phylogenies. It also allows for parsimony reconstruction and bootstrapping.

PHYLIP: `http://evolution.gs.washington.edu/phylip/`. This phylogeny program, written by Joe Felsenstein, is one of the oldest and most popular maximum likelihood software packages. Supported data types include DNA sequences, RNA sequences, protein sequences, discrete characters, and continuous characters (i.e. gene frequencies). This package also includes programs to carry out parsimony analysis and distance matrix methods.