7

PARAMETRIC PHYLOGENETICS

> Given that we select estimated evolutionary trees always according to the maximum likelihood criterion, the method for constructing an estimated evolutionary tree is in principle well determined once a stochastic model of the evolutionary process has been selected.
>
> —James S. Farris, 1973

Parsimony analysis is only one of several methods for analyzing phylogenetic relationships. In this chapter, we examine two other approaches: maximum likelihood (ML) and Bayesian analyses. Huelsenbeck and Crandall (1997) provide a general introduction to likelihood analysis. Lewis (2001a) provides a general introduction to Bayesian analysis, and Holder and Lewis (2003) provide a useful introduction to several approaches including ML and Bayesian and parsimony analyses. ML and Bayesian analysis are similar in using likelihood calculations as the basis for inference, but they are very different in their philosophical approach to problem solving. ML methods use a criterion-based approach: the preferred tree is the tree that has the highest probability of producing the data we observe given a specific model of evolution adopted by the investigator, the tree topology, and the branch lengths between nodes. The model is used to calculate probabilities of observing the data on a specified tree, one transformation series at a time. It differs from parsimony in taking branch length into consideration relative to an explicit model of change from one character state to another along the branch. Specifically, the longer the branch, the lower the probability that matches (e.g., identical base residues) appearing at the base of the branch and the tip of the branch are homologous. These probabilities

Phylogenetics: Theory and Practice of Phylogenetic Systematics, Second Edition.
E. O. Wiley and Bruce S. Lieberman.

are then multiplied over the entire tree to produce a likelihood score. This is often expressed as a logarithm of the likelihood for purely computational reasons. When estimating the topology, the tree with the highest log-likelihood scores is accepted as the best estimate. Like parsimony, likelihood produces a point estimate of the best tree given the criterion such that the fitted model (data fitted on a tree topology and other parameters) is the "best model."

Bayesian analysis uses likelihood calculations, but stands the probabilities on their head by estimating the probability of the tree topology given the data and the model rather than the probability of the data given the model and tree topology. It does so by calculating a posterior probability, a probability that is conditional on what the investigator is willing to accept as true before the analysis. Most frequently, this prior probability states that all alternative trees are equally probable, a priori, but it is entirely possible to specify a particular tree topology as the most probable. Bayesian inference offers an entirely different approach to statistical inference and is controversial among statisticians, as we shall discuss below. As practiced by phylogeneticists, in Bayesian inference the criterion employed is that of maximizing the posterior probability of the tree given the data and model of inference. It uses algorithms that are designed to explore probability space to find the area(s) where the probability density is highest and, thus, expected to be the area where the tree(s) with the highest posterior probability reside. Unlike either likelihood or parsimony, Bayesian analysis does not produce a point estimate of the model, but rather, a probability distribution of models that may contain one to many tree topologies.

The use of explicit models of evolution to reconstruct phylogenetic relationships and accompanying statistical tests has increased tremendously in the past 20 years, due in part to the increased availability of computer packages capable of performing such analyses (e.g., PHYLIP, PAUP, MrBayes) and the application of the method of Felsentsein (1981b) in these programs that dramatically decreased the computational burden of the calculations. However, the idea that likelihood models could be used for phylogenetic inference dates back to at least Cavalli-Sforza and Edwards (1964).

A phylogenetic researcher might be motivated to incorporate evolutionary models for a number of reasons. For instance, investigators may believe that they understand the evolutionary dynamics of some kinds of data well enough to model their expected evolution over a particular phylogeny. This a priori belief is no different in kind than that of an investigator who weights the performance of certain characters in a parsimony analysis or even the a priori choice of one kind of character over another; any such activities may or may not be warranted. Both investigators are motivated to increase performance of the analysis, increasing what they believe is the veracity of the resulting phylogenetic hypothesis, by applying their experience to the problem. The outcome, for better or worse, is left to other investigators and subsequent researchers to evaluate. If the evolutionary dynamics of the model capture something real about the data, the veracity of the result is better. If led astray, the result may be worse.

The growth of this particular branch of phylogenetics is likely to continue at an almost exponential rate. In this chapter, we will approach statistical phylogenetics in a narrative fashion. We will cover the basics of likelihood-based approaches with simple examples and concentrate on how they work, keeping mathematical formalities to a minimum. The goal is to understand how such approaches work and why we should choose, or not choose, to use them. We leave the details to others who are better qualified to describe the mathematical formalities in detail (e.g., Swofford et al., 1996; Felsenstein, 2004; Yang, 2006). We also recommend Sober's (2008) account of the philosophical basis for inference as it relates directly to issues of likelihood, Bayesian, and parsimony inference.

MAXIMUM LIKELIHOOD TECHNIQUES

Likelihood is the probability that an event that has happened in the past would yield a specific outcome. ML is a procedure for finding the value of one or more parameters for a given model that makes the likelihood of observing the data at hand attain its maximum value. Consider a very simple example of some data taken from a population of fishes, the length and height of the body. We have adapted this example from a similar example given in Forester and Sober (1994) and will return to it later in the chapter. We might attempt to fit a line to these data to investigate how they are related. Two such lines are shown in Fig. 7.1. It should be obvious that the bottom line fits the observations better than the top line. If we consider both lines to be "models," then we can intuit that the data are better explained by the bottom line than the top line and, thus, the bottom line would have a higher likelihood value than the top line. Indeed, we can fit any straight line to these data, calculate a likelihood value under the assumption of a normal distribution of errors, and compare it to our lower line. Some will be a very bad fit, but some will be a fairly good fit.

As revealed by this very simple example, the basic idea of ML is quite simple: the best line is the line that is most consistent with the observations. It "predicts" (postpredicts or "postdicts") the data better than other lines. Consistency is measured statistically, by the probability that the observations should have been made if the line-model is taken as true. We can adjust the model (including its parameters) and test whether a new line is better than the old line. If the likelihood goes up,

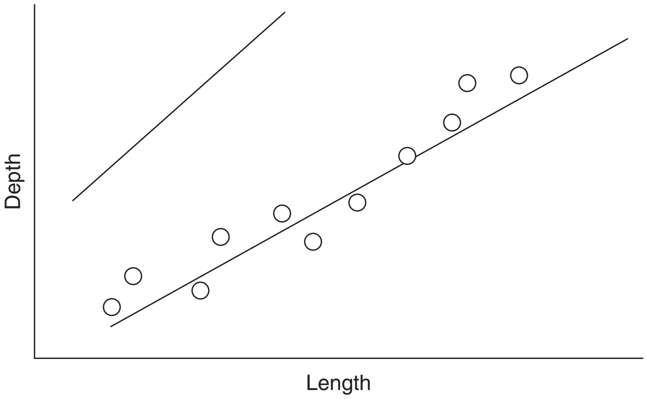

Figure 7.1. A scatter plot with two regression lines. Intuitively we can guess that the lower line fits the data (circles) better than the top line.

then the new line predicts the data better than the old line. We can continue to adjust the parameter values until such time as we fail to obtain a higher likelihood value. Or we can adopt an entirely new form of model, for example, a curved line rather than a straight line. It is actually more complicated, but this will get us started.

You might ask: why not simply calculate the regression line using the least squares method? Of course, you could do this and you probably would, given this simple example. If you simply calculated the regression, you would be using the algorithmic approach rather than the criterion-driven approach. Algorithmic methods work well in many areas of statistical analysis, and the algorithmic approach is preferred because it arrives at the ML solution in a faster and more efficient manner (just as algorithmic methods are faster than criterion-based methods in parsimony). Indeed, the regression line is the ML estimator. But phylogeny reconstruction is harder and the parallels of algorithmic versus criterion-driven approaches discussed in the last chapter are entirely applicable to likelihood-based analyses of phylogenetic problems.

Most explanations of ML methods begin with a demonstration of coin flipping. Read (2000) uses a very simple example that we believe illustrates the process in a way that can be followed through to more complex as well as Bayesian examples. Consider a population of organisms. The data consists of antennae lengths. We sampled 20 individuals. What we need now is a model. Where we get the model depends on our past experience. In this case, our experience leads us to assume that populations of measurements such as these have some unimodal distribution, central tendencies, and that the observations are expected to deviate from the central tendency in some predictable fashion given certain assumptions about variation. Thus, we will expect the population of measures at hand have some mean value and a standard deviation (SD), which is a measure of the scatter of the data points relative to the mean. The simplest case is to consider the central tendency of the population from which the samples were drawn is μ and the variation around this tendency is σ^2. If we assume that the data are distributed according to a normal (or Gaussian) distribution, then we can calculate the likelihood of the data given the model by calculating the probabilities of each piece of data for any chosen values for the mean and standard deviation. We can compare the fit of the data to the model and its associated parameters.

Consider Fig. 7.2a. This figure is a plot of the normal curve with a mean antenna length of 5 mm and a standard deviation of 1. The x-axis is the antenna length, and the y-axis is the probability density associated with finding a particular value of antenna length. We will tell you now that these parameter values are the true parameter values for this population. Note that the highest probability densities [p(x)] for observing particular antennae lengths are associated with the value of 5 mm and that the probability density decreases as antenna length increases or decreases from the mean value. Another way of saying the same thing is to say that our probability density is high (more measures have a higher probability of being observed). This is intuitively understandable; we expect to observe values closer to the true mean more frequently than values farther away from the true mean. In other words, we have the expectation that the probability of a measure close to the mean is larger than the probability of a measure farther from the mean. We have also labeled the particular lengths of two specimens. The line from the specimen value to the curve is a measure of the probability density of encountering the particular measurement

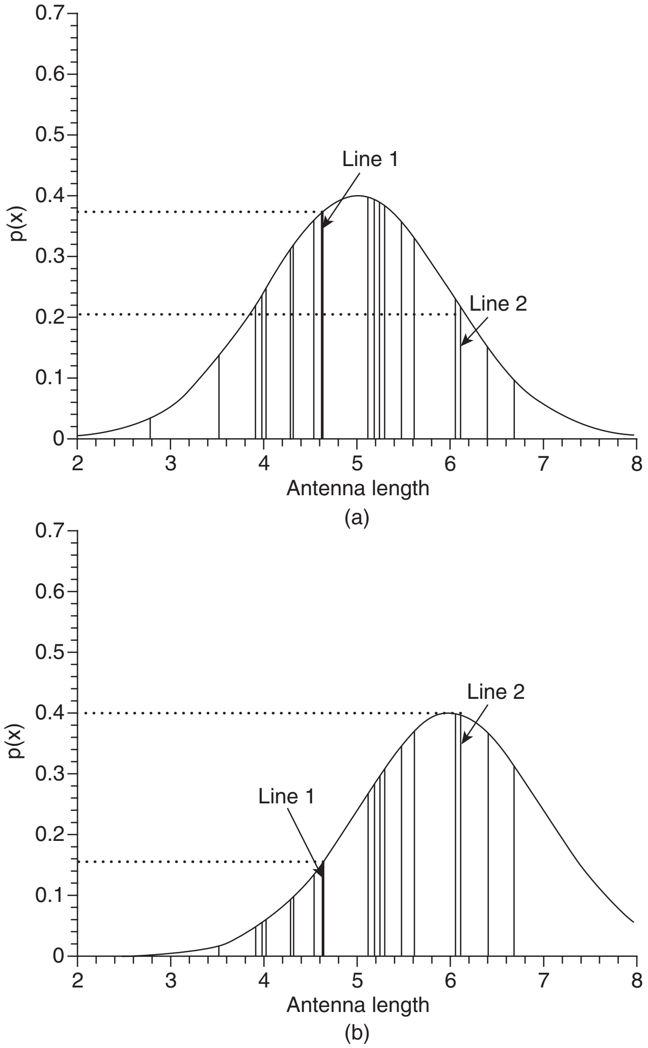

Figure 7.2. Two models of the mean and standard deviation of the antenna lengths for a species of insect. X-axis is antenna length. Y-axis is the probability of observing a measure. (a) Model with mean = 5 mm, standard deviation = 1 mm. (b) Mean = 6 mm, standard deviation = 1 mm. Each line drawn from the x-axis is a measure drawn to the length of its probability of occurrence in the population. The dotted lines are drawn from selected measures to the probability axis. Modified from Read (2000), used with permission.

given the model parameters (i.e., given that mean = 5 and standard deviation = 1). We can read off the values of each measure on the probability axis.

Now consider Fig. 7.2b. The same plots for the same measured antennae are shown. We have kept the same general model, but we have changed the model parameter values. In this case, the model has a mean of 6 mm and a standard deviation of 1. Note that Line 1 is shorter, indicating that the probability of encountering this measurement is less in this model than in the model shown in Fig. 7.2a. Note

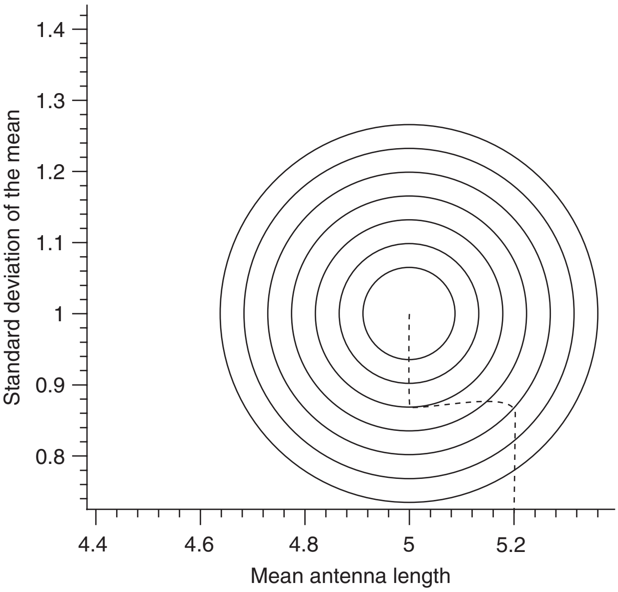

Figure 7.3. A contour map of the probability density space in mean and standard deviation for the example in Fig. 7.2. The dotted line is an imaginary route taken by fitting different values to a fitted model from less probable to more probable. Modifed from Read (2000), used with permission.

that Line 2 is longer indicating that the probability of encountering this value is greater if the mean of the population is 6 mm and the standard deviation is 1. Now, consider all 20 measurements and the two models. If you took the time to actually measure the lengths of each line, you would find that encountering these 20 measurements is more probable if the mean = 5 mm than if the mean = 6 mm, because the total length of all lines together is greater when the mean = 5 than when the mean = 6. The same is also true if you compare a model with mean = 5 and standard deviation = 1 with a model where mean = 5 and standard deviation = 0.6 or 2.0. Thus, we conclude that a fitted model with the parameter values mean = 5, standard deviation = 1 maximized the likelihood of observing the 20 antenna lengths. Although we have visualized these parameters as simple graphs, we can also visualize the relationship between the standard deviation and the mean as a contour map (Fig. 7.3). Such a map could be prepared by plotting the results of many likelihood models. It permits a look at the models relative to the data as compared to the data relative to the models.

The connection between ML and simple statistics in this example is transparent. The ML of the data corresponds to a fitted model where the mean and standard deviation of the sampled lengths follow a normal distribution and have values of mean = 5 mm and standard deviation = 1.0 (the standard deviation being biased relative to the population standard deviation).

In this simple example, the choice of using this particular model may seem unproblematic. But there are deeper issues. Models are simplifications, which means that in all cases none of the models one is evaluating is a true model. Sober (2008) makes the point that the value of models is not whether they are true but whether they are good at predicting in the face of new data. If they are successful in predicting new data, then the fit of data to the model may lead to something approximating truth in nature. In this example, which is a simple problem of estimating means, the fitted model predicts that the next draw of samples will have estimated means and standard deviations within a range predicted by the fitted model mean standard deviation. If this prediction is fulfilled, then the fitted model will have led to a valid estimate of something that exists in nature.

Simplicity

This example leads us to an interesting feature of models and their parameters. It is embedded in philosophical controversies surrounding simplicity and parsimony. Philosophers have often wondered why simple explanations are to be preferred over more complex explanations (e.g., Sober, 1975; Forster and Sober, 1994; Sober, 2008). There are actually two questions. Given a model and some parameters, which values assigned to the parameters do the best job of explaining the data? This one is fairly easy. In Fig. 7.1, we see that our best-fitting regression line explains the observed data better. In parsimony analysis under a model that all changes are equally weighted, the shortest tree is to be preferred over a longer tree because it describes the data better (Farris, 1983). However, there is also the question of which parameters to choose for the model. Do we choose the smallest number of parameters that adequately "explain" the data? In the example involving means (Fig. 7.2 and discussion above), there are several kinds of models. We can distinguish them by their parameters. The first model contains the mean and standard deviation; a second contains mean, standard deviation, and skewness. The models are nested, because the model containing the parameters mean and standard deviation are included in the model containing mean, standard deviation, and skewness. In this case, it is a simple matter to determine if there is a significant difference in likelihood between the models. A similar scenario can be applied to phylogenetic analysis. Does adding a new parameter result in a significantly better fit to the data? If so, add the parameter. If not, the parameter is not contributing to the resulting estimate in any significant manner.

Akaike (1973) suggests that we can determine the need to add extra parameters in explaining the data at hand by taking into account the relative "burden" of these extra parameters in explaining the data. Given the data in our example, 20 measurements, and that the population is slightly skewed in its observed distribution of data values, we can intuitively (and statistically) see that adding a parameter for skewness does not significantly improve the fit. Akaike (1973) provides a formal way of determining the burden of adding extra parameters that takes into account the simplicity of models as well as their ability to fit the data. Forster and Sober (1994) provide an accessible description of these formalities.

In essence, here is the problem. Even if a model is true, this does not result in perfect data because observations contain error. Return to our first example of fitting a line to a series of points (see Fig. 7.4): Consider that the line is true (that is, that the model is true) and that the points were generated from this line, rather than the line being generated to fit the points. Because the points do not fall exactly

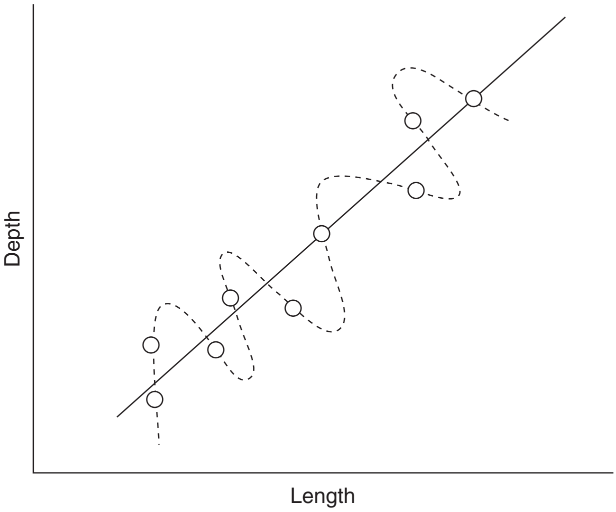

Figure 7.4. Graphic representation of two models that explain the relationship between length and depth of a hypothetical population with circles indicating actual measures. The straight line represents a fitted regression model that is most predictive of new data that might be collected. The dotted line is chasing error.

on the line, we can assume that the reason they do not is due to sample variation or measurement error. Now, we could fit a very complex line to these data, "explaining" all of the points (Fig. 7.4, dotted line). This line would contain a great number of parameters. But the extra parameters would actually be explaining the error, missing the truth. It would be "over parameterized." Akaike's criterion contains within its framework a mechanism for addressing this problem because it takes into consideration both the fit of the data and the simplicity of the model. The more complex the model, the more it must explain if it is to be the preferred model.

There is another thing we can learn from the example of slightly skewed data. Remember that we were using only 20 observations of antennae length. Recall that skewness does not help because the difference in likelihood found with and without considering skewness was not significant. But what if we had 10,000 measurements? It is quite possible that we would then need that extra parameter. This makes sense because our judgments are made relative to data, and not in a vacuum.

We can think of the model selection process as that area of parametric tree building where parsimony plays a part. Parameter-rich models can lead to the inability to discriminate between trees (Yang et al., 1995). The simplest model that can be implemented is preferred over more complex models.

Likelihood in Phylogenetics: An Intuitive Introduction

We introduced an intuitive sense of likelihood using a single correlation example (Fig. 7.1). Given what we have learned of parsimony and parsimony analysis, we

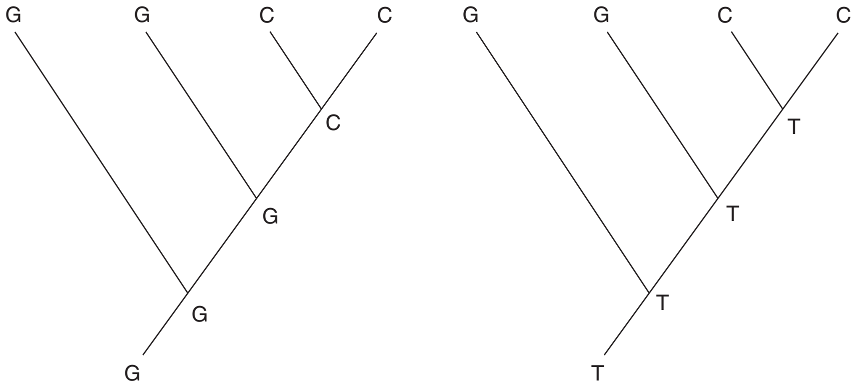

Figure 7.5. Two trees with identical topologies of bases assigned to the leafs but with different bases assigned to the nodes. Intuitively, the assignment of bases on the left tree has a higher probability than those on the right tree given either the parsimony or likelihood criterion.

can gain some intuitive sense of likelihood by considering the two trees in Fig. 7.5. These trees are identical in topology and observed data (shown at the tips) but are assigned different bases at the nodes. Note the distribution of the bases in the terminal taxa and the assignment of bases in the nodes. Which assignment of bases at the nodes do you think are more probable? Intuitively, we might agree that the assignment of bases on the left is more probable, given the data actually observed in the terminals. Parsimony gives us this same conclusion. The tree on the left requires but one change (G to C) while the one on the right requires four changes. As a parsimony problem, we would say that the character state assignments at nodes are more optimal in the tree on the left because the assignments on the left-hand tree require fewer steps. It is no wonder that we reach both conclusions: Sober (1983a, b, 1987) has argued that likelihood and parsimony converge to the same result if characters are considered one at a time under a simple model of character state change; Tuffley and Steel (1997) proved this result given certain conditions such as a Junks-Cantor-like model and different sets of branch lengths for each character.

However, we might ask another question: would the assignment of bases on the right tree be impossible? Answer: no, it is simply highly improbable given the topology and the distribution of bases on the terminals. If we were using parsimony, we might consider adenine to be allocated to the basal node, but we would never consider allocating adenine to the node leading to two terminal cytosines. We might concede that this is possible, but our character optimization algorithms do not allow for the result. Using likelihood, however, we can calculate a likelihood score for the distribution of bases on the right-hand tree as well as the bases on the left-hand tree, with a model of change and some math. In fact, we can calculate all of the various possible transformations that might occur no matter how improbable they may be if we have such a model. There would be 16 such possible base assignments for this particular problem of 4 terminals and 3 internal nodes, each specifying a particular scenario and each being relatively more or less likely, given the model. If we added all of these calculations up, we could calculate a total likelihood for the entire tree over all possible transformations for this character. This is what a likelihood analysis accomplishes for any particular tree topology; it sums all of the likelihoods for all possible evolutionary scenarios of each site. It returns a value, that is, a sum of the log likelihoods for characters. The "best tree" is the one that maximizes this value.

As it turns out, we can do the same thing for parsimony. We could run all 16 permutations and evaluate the results in terms of tree length. The essential differences are that likelihood embodies a model of change that takes into account the length of the branch connecting two nodes and in likelihood we typically sum over all possible ancestral character states while in parsimony the score is based on the evolutionary pattern that yields the shortest length. Consider the C at the node leading to the two terminals (Fig. 7.5). The model specifies the probabilities of C remaining C or changing to A, T, or G over time or a surrogate of time, branch length. For relatively short branch lengths (definition: relatively short periods of time or relatively slow rates of mutation and fixation), the probability that a C in the terminal had an ancestral C at the node is relatively high, perhaps approaching 1. But over relatively long branch lengths, it becomes increasingly probable that the C at the node will have changed to some other base(s), which then changed back to C. This is under the evolutionary assumption that changes from C to some other base are governed by the model adopted.

The total likelihood for one tree can be compared to likelihood values obtained for other trees just as tree length values can be compared between parsimony trees. The tree with the highest likelihood score is preferred over alternative topologies because it fulfills the optimality criterion (ML) better than its competitors.

Likelihood in Phylogenetics: A More Formal Introduction

ML analyses in phylogenetics evaluate the likelihood of observing the data on a particular tree topology, given a particular model of evolution. If we consider phylogenetic analysis alone, we are searching for the best topology. Among the alternative topologies, that topology with a higher probability of giving rise to the observed data is preferred to topologies with lower probabilities, given the model adopted (Swofford et al., 1996). In certain respects, this is similar to how parsimony works. For instance, given an array of possible tree topologies, how well do the characters fit that topology relative to a specified criterion and the assumptions of the analysis? In parsimony, as mentioned previously, the criterion is minimum number of steps and the assumptions have been detailed in Chapter 6.

Likelihood is calculated for a given tree topology and a given set of branch lengths. Unlike parsimony, every possible transformation in each character is calculated over the topology and the branch lengths of that topology. Each topology has a potentially infinite number of associated branch lengths. An example, provided by Swofford et al. (1996) and redrawn in Fig. 7.6, illustrates the process. Note that while the most likely character state at node 5 is "C," the likelihood of all four bases at that node and all possible derivations of the C from the node below it also are calculated. Given a particular tree topology, the basic steps in performing an ML analysis are listed below.

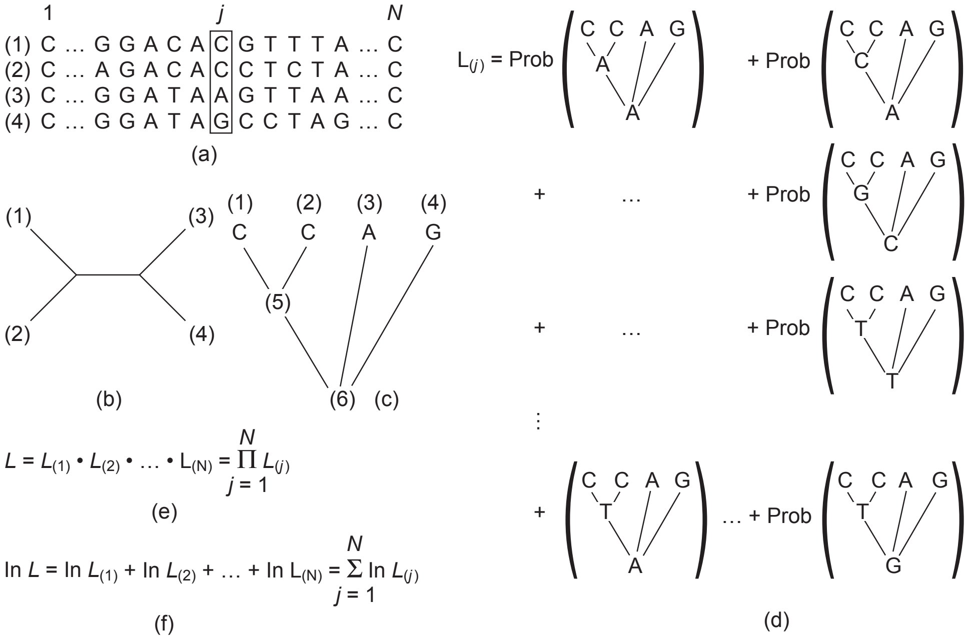
1 j N
(1) C ... G G A C A C G T T T A ... C
(2) C ... A G A C A C C T C T A ... C
(3) C ... G G A T A A G T T A A ... C
(4) C ... G G A T A G C C T A G ... C
(a)
(1) (3) (2) (4)
(b)
(1) C (2) C (3) A (4) G
(5)
(6) (c)
$L = L_{(1)} \bullet L_{(2)} \bullet \ldots \bullet L_{(N)} = \prod_{j=1}^{N} L_{(j)}$
(e)
$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}$
(f)
L(j) = Prob
+ Prob
+ ...
+ Prob
C C A G
A
C
G
T
... + Prob
(d)

Figure 7.6. The flow of likelihood calculations. (a) A data matrix of four taxa (1–4) and DNA bases (ATCG); we will use the data in column "j" for the example. (b) An unrooted tree with the characters from column "j." (c) A rooted tree with the same data; we will follow the process using only this tree, but a full account would require us to calculate likelihoods for other trees and compare them in some fashion. (d) Sample of the 16 various assignments that could be made to the nodes of tree (c); the sum of the likelihoods of all 16 constitute one of the elements of the next two figures (i.e., L_j would be the sum of the 16 possibilities for data column "j"). (e) The all "N" columns of data $[L_{(1)} \ldots L_{(N)}]$, the likelihood of the tree (c) is the sum of the likelihoods for all data columns. (f) The usually reported value is the log likelihood, which is the sum of the log likelihoods of each data column. Copyright Sinauer Associates, Inc., used with permission.

1. For each character of the data matrix, calculate the likelihood of character transformation at each node of a topology and its associated branch lengths, from the tips to the root. This is accomplished using the evolutionary model in concert with the observed character distributions to calculate the probabilities of observing character states at nodes.
2. Calculate the likelihood of the entire tree by multiplying the probabilities of each character together under the assumption that each character is evolving independently. That is, the likelihood of the entire tree is the product of the likelihoods of each character (each site) and this is usually expressed by summing the log of the likelihoods (because the product of likelihoods is very small).
3. Among the trees evaluated, pick that tree with the highest likelihood.

Note that the likelihood approach is a contrastive approach (Sober, 2008). Likelihood scores have no direct interpretation unlike the case in parsimony where steps have a direct interpretation. It is the difference between two likelihood scores that is important, not the scores themselves.

There are a variety of models that might be used to calculate the probabilities for character state change. For molecular data, these models are expressed in a matrix of change that may be overlain with assumptions about variation in rates over characters. Consider the following matrix of possible changes from one base to another:

$$\begin{array}{ccccc}
 & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\
\mathrm{A} & \mathrm{A}\to\mathrm{A} & \mathrm{A}\to\mathrm{C} & \mathrm{A}\to\mathrm{G} & \mathrm{A}\to\mathrm{T} \\
\mathrm{C} & \mathrm{C}\to\mathrm{A} & \mathrm{C}\to\mathrm{C} & \mathrm{C}\to\mathrm{G} & \mathrm{C}\to\mathrm{T} \\
\mathrm{G} & \mathrm{G}\to\mathrm{A} & \mathrm{G}\to\mathrm{C} & \mathrm{G}\to\mathrm{G} & \mathrm{G}\to\mathrm{T} \\
\mathrm{T} & \mathrm{T}\to\mathrm{A} & \mathrm{T}\to\mathrm{C} & \mathrm{T}\to\mathrm{G} & \mathrm{T}\to\mathrm{T}
\end{array}$$

Consider → to specify some model of change from one base to the other (and the probability of not changing). We can scale the rates such that the branch lengths are in terms of the expected number of substitutions per site (column of data, character). The rate of "leaving" a state is reflected in the diagonal of the rate matrix, as in A→A or C→C. This rate is simply a negative number with a magnitude equal to the sum of the rates of changing to each of the other states, e.g. A→C or A→G. So, A→A is: –(A→C+A→G+A→T).

$$\begin{array}{ccccc}
 & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\
\mathrm{A} & -(\mathrm{A}\to\mathrm{C}+\mathrm{A}\to\mathrm{G}+\mathrm{A}\to\mathrm{T}) & \mathrm{A}\to\mathrm{C} & \mathrm{A}\to\mathrm{G} & \mathrm{A}\to\mathrm{T} \\
\mathrm{C} & \mathrm{C}\to\mathrm{A} & -(\mathrm{C}\to\mathrm{A}+\mathrm{C}\to\mathrm{G}+\mathrm{C}\to\mathrm{T}) & \mathrm{C}\to\mathrm{G} & \mathrm{C}\to\mathrm{T} \\
\mathrm{G} & \mathrm{G}\to\mathrm{A} & \mathrm{G}\to\mathrm{C} & -(\mathrm{G}\to\mathrm{A}+\mathrm{G}\to\mathrm{C}+\mathrm{G}\to\mathrm{T}) & \mathrm{G}\to\mathrm{T} \\
\mathrm{T} & \mathrm{T}\to\mathrm{A} & \mathrm{T}\to\mathrm{C} & \mathrm{T}\to\mathrm{G} & -(\mathrm{T}\to\mathrm{A}+\mathrm{T}\to\mathrm{C}+\mathrm{T}\to\mathrm{G})
\end{array}$$

Now consider one cell in the matrix: A→C. There are at least three components to the transition rate from A to C along any one branch of the tree: the relative frequency of base C in the model, the relative rate parameter for the A→C transformation, and the instantaneous substitution parameter. The second component, the relative rate, asks a question: how often does A change to C relative to a change of A to G or a change of T to C? We can write this relative rate into a matrix, discounting bases changing into themselves, as shown below:

$$\begin{array}{ccccc}
 & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\
\mathrm{A} & — & \mathrm{A} & \mathrm{B} & \mathrm{C} \\
\mathrm{C} & \mathrm{G} & — & \mathrm{D} & \mathrm{E} \\
\mathrm{G} & \mathrm{H} & \mathrm{I} & — & \mathrm{F} \\
\mathrm{T} & \mathrm{J} & \mathrm{K} & \mathrm{L} & —
\end{array}$$

If we consider evolution to be time-reversible, as we usually do in parsimony calculations, then the relative rate of change of A to C would be the same as that from C to A, and we can simplify the matrix, because A = G, B = H, etc., as shown below:

	A	C	G	T
A	—	A	B	C
C	A	—	D	E
G	B	D	—	F
T	C	E	F	—

Let us consider this component of the substitution matrix. The values of the relative rate parameter depend on the assumptions of the evolutionary model. The simplest model is Jukes-Cantor (1969). This model assumes that there is an equal probability of any base changing into any other base. In short, the relative rates are equal. So the matrix is quite simple:

	A	C	G	T
A	—	A = 1	B = 1	C = 1
C	A = 1	—	D = 1	E = 1
G	B = 1	D = 1	—	F = 1
T	C = 1	E = 1	F = 1	—

Another relatively simple model, Kimura (1980), allows that there are differences between the rates between transitions and transversions. Thus, the model is a bit more complicated. For example, in the matrix shown below, κ determines how much more likely transversions are than transitions. If we set $\kappa = 2$, then the rate of a transition is twice the rate of a transversion.

	A	C	G	T
A	—	1	κ	1
C	1	—	1	κ
G	κ	1	—	1
T	1	κ	1	—

The next component, the instantaneous rate component (μ) is the probability that a change will occur at some very short time period along a branch. In time-reversible models, we also have to account for how common the base is assumed to be, because the actual rate of change is a function of the frequency of the base. Thus, for any one transformation, we have to account for three factors, not two:

$$A \rightarrow C = \mu a \Pi_C, \text{ given symmetrical change, or}$$

$$A \rightarrow A = -(\mu a \Pi_{C} + \mu b \Pi_{G} + \mu c \Pi_T) = -\mu(a\Pi_{C} + b\Pi_{G} + c\Pi_T)$$

We can sort all this out into a very general matrix, termed the Q-matrix. It can be described as a matrix that models the rate of change from one base to another during some very small amount of time. For a time-reversible model:

	A	C	G	T
A	$-\mu(a\Pi_C + b\Pi_G + c\Pi_T)$	$\mu a\Pi_C$	$\mu b\Pi_G$	$\mu c\Pi_T$
C	$\mu a\Pi_A$	$-\mu(a\Pi_A + d\Pi_G + e\Pi_T)$	$\mu d\Pi_G$	$\mu e\Pi_T$
G	$\mu b\Pi_A$	$\mu d\Pi_C$	$-\mu(b\Pi_A + d\Pi_C + f\Pi_T)$	$\mu f\Pi_T$
T	$\mu c\Pi_A$	$\mu e\Pi_C$	$\mu f\Pi_G$	$-\mu(c\Pi_A + e\Pi_C + f\Pi_G)$

Note that two of the factors are related as rate parameters while the other is a frequency parameter. For any small amount of time, the actual rate of substitution for A→C is a function of both the relative rate parameter and the instantaneous substitution rate, or μa, coupled with the frequency of the base Π_{base}. As it turns out, the Q-matrix is the product of two matrices (i.e., it can be decomposed into two matrixes), R and Π:

$$R=\begin{matrix} — & \mu a & \mu b & \mu c \\ \mu g & — & \mu d & \mu e \\ \mu h & \mu j & — & \mu f \\ \mu i & \mu k & \mu l & — \end{matrix} \qquad \Pi=\begin{matrix} \Pi_A & 0 & 0 & 0 \\ 0 & \Pi_C & 0 & 0 \\ 0 & 0 & \Pi_G & 0 \\ 0 & 0 & 0 & \Pi_T \end{matrix}$$

Once we have a model, we can use it to calculate changes from one base to another along any particular branch of a tree over time by calculating a transition (or substitution) probability:

$$P(t) = e^{Qt}$$

The transition probability matrix can be evaluated for any particular branch length by decomposing the Q-matrix into it eigenvalues and eigenvectors. Swofford et al. (1996) demonstrate how this is accomplished for some of the simpler models where relatively simple expressions exist for the eigenvalues. For example, the Jukes-Cantor model only states that a particular base can change or not change, so there are only two probabilities to consider where i and j denote bases:

$$P_{ij}(t) = \tfrac{1}{4} + (\tfrac{3}{4})e^{-\mu t} \text{ if } i = j \text{ (no change)}$$

$$P_{ij}(t) = \tfrac{1}{4} - (\tfrac{1}{4})e^{-\mu t} \text{ if } i \neq j \text{ (change)}$$

We considered parameters earlier. Now that we know what a substitution model looks like, we can relate the concept of parameters to phylogenetic likelihood analysis. If a model has a parameter that receives a constant value over the entire model, then that parameter is effectively factored out. For example, if we assume the Jukes-Cantor model, then a = b = c = d = e, etc. Further, Jukes-Cantor assumes equal base frequencies: $\Pi_A = \Pi_B = \Pi_c = \Pi_D$, and this parameter factors out. Thus, we are left only with a single parameter, μ, and a single parameter model. You can see this in the equations above, it all boils down to a function of μ. Kimura (1980) introduces two values for the relative rate parameter, but assumes that base frequencies are equal; so here we have a two-parameter model. We can continue to add parameters, for example, the eight parameters of the general time reversible (GTR: Tavaré, 1986) model.

Most likelihood analyses use a series of assumptions built into the analysis that simplify calculations.

1. Most models are time-reversible. This permits one to calculate likelihoods by rooting the tree at an arbitrary node and decreases computational effort. The result is an unrooted tree.

2. Characters (nucleotide sites, base positions) are assumed to evolve independently (Kluge's auxiliary principle). This permits likelihoods to be calculated separately for each character and then multiplied in order to calculate a likelihood for the entire tree.
3. Change is assumed to follow a Markov process. We assume that changes along different branches of the tree are independent of each other. In addition, the rate of particular change, e.g., A to T, does not depend on the history of change of the site prior to the acquisition of A.
4. Change is usually assumed to follow a homogeneous Markov process. Homogeneous Markov processes assume that the rate of change from one particular character state to another state (as specified by the evolutionary model) is the same over the entire tree.

Simple models assume equal base frequencies. However, the frequency parameter is often set empirically, by calculating the relative percentages of the four bases in the data matrix. The relative frequency of each base is assumed to be constant over the entire tree. The product of the relative rate parameter and the instantaneous substitution parameter yield a rate parameter that specified, in essence, base turnover rate along branches.

The probability of changing from one base to another is a function of the substitution rate and time. The mean substitution rate is set to one, and the relative rates are scaled so that they sum to the mean rate (Yang, 1994). The rate of evolution is usually allowed to vary over the tree. In this form of inference, each branch on the tree has a parameter that represents its length in terms of the expected number of changes per character. Forcing the rate to be fixed over the tree is adopting the molecular clock hypothesis, and likelihood can be calculated by considering times of divergence in rooted trees (Swofford et al., 1996, and citations).

Calculating the likelihood of a particular tree requires us to consider the likelihoods of the occurrence of each character state at each node given the states of the terminal taxa, the tree topology, and the estimated branch lengths. For the data at hand and a specified tree, likelihoods are calculated taking into account the prior probability of a particular character state (based, for example, on its overall frequency in the data) and the conditional probabilities of the character state remaining the same or changing along branches from the root to terminal taxa (observed sequences). Summing all this leads to a likelihood for the tree as a whole. To find the ML value for a tree, many combinations of parameter values must be evaluated until we find a set of values for all parameters (including branch lengths) that maximizes the likelihood.

The substitution probability matrix can be modified to accommodate rate heterogeneity by adding a rate factor based on some a priori or empirical assessment of variation among sites and rate of change. In fact, if there is significant rate variation among sites and this factor is left out of the model, then likelihood analyses will suffer from some of the faults typically ascribed to parsimony such as "long branch attraction" (Gaut and Lewis, 1995). A simple discrete model is the invariable-sites model (Hasegawa et al., 1985), where some fraction of the sites is considered invariant while the remaining fraction vary at the same rate. The gamma distribution

(Yang, 1993; Steel et al., 1993) provides a continuous model of rate variation among characters, although the model is usually implemented as an approximation that uses discrete rate categories (Yang, 1994). Typically we set the β parameter of the Gamma distribution 1/α, so that the mean rate across all characters is 1.0 and low values of α denote distributions of variation among sites such that most sites are evolving slowly while a few are evolving at a rapid rate. Higher α-values lead to less heterogeneity in rate, and if α is infinite, then all sites are evolving at exactly the same rate. The value of α can be inferred using ML, as with the other parameters of the model.

Selecting Models

Likelihood methods present the investigator with a plethora of models. The question is: what model should one use? One problem in determining this is that different models contain different numbers of parameters, and as the number of parameters increases, so does the variance associated with the ML values of a given set of trees. We can fall into the trap of tracking noise rather than signal. All parameters except the one the investigator is interested in are termed "nuisance parameters." If the parameter you are seeking is the tree topology, then all other parameters such as branch lengths or rates of change in particular classes of data are "nuisance parameters." However, as the investigator introduces more and more parameters, the possibility arises that the new parameters are contributing little to the result, which is the tree topology. Consider our original example of estimating the mean. We added a parameter of skewness, but our data were hardly skewed and adding the parameter did not significantly contribute to our estimate of the mean of the population. So parameters should be added only when they actually help and the process of adding parameters can reach a point in model complexity where one cannot discriminate between any of the trees due to the large variances associated with each parameter. Thus a model is said to be "over-parameterized." Over parameterization is a well-known statistical problem in such operations as multiple regression and discriminant analysis.

One may wonder how an investigator might pick a model in the absence of a known phylogeny. Swofford et al. (1996) suggest that one selects a goodness-of-fit statistic and then selects a model that maximizes this statistic without adding additional (perhaps unnecessary) parameters. Many such tests or criteria can be used to inform the investigator of the fact that adding an additional parameter does not significantly improve the fit. The first is the log likelihood ratio test (the G-test of Sokal and Roth, 1981), and the second is the Akaike information criterion (Akaike, 1973; Sober, 2008).

The log likelihood ratio test is simply the likelihood ratio test of goodness of fit. It tests the significance between two nested hypotheses (that is, one hypothesis is a proper subset of the other, as in our example of mean and standard deviation versus mean, standard deviation, and skewness). In phylogenetics, for example, we can select a particular tree topology and then compare models with different numbers of parameters. In application, one sees if adding an additional parameter to a model produces a significant increase in goodness-of-fit of the data *given the same tree topology.* For example, consider a model of equal rates of evolution across all sites and another that allows rates to vary. The models are nested and can be tested against a common topology. Does the more complicated model (varying rates) produce a significantly better fit of the data than the model of equal rates across all sites? If so, then one accepts the more complicated model.

You might wonder how the topology is picked. Sullivan et al. (1997) showed that using a topology that is a rough estimate of the actual phylogeny will result in a reasonable model for further analysis. A good rough estimate is a parsimony tree (Yang et al. 1995; Sullivan and Swofford, 2001). However, a random topology or one very dissimilar to the actual phylogeny can produce suboptimal models (Sullivan et al., 1997).

The Akaike criterion introduced a penalty for each added parameter based on the degrees of freedom associated with the parameters (Akaike, 1973). Its principal strength is that it can be used to test across tree topologies, rather than being restricted to testing within nested hypotheses (Prosada and Buckley, 2004; Sullivan and Joyce, 2005).

BAYESIAN ANALYSIS

Adoption of likelihood approaches to phylogenetic systematics lead to consideration of Bayesian approaches. Likelihood asks "what is the probability of the data given the model [p(data|model)]? The alternative question is: what is the probability that the model is correct given the data [p(model|data)]? This has to be evaluated with respect to a set of candidate models. The relationships between these two questions are bridged by Bayes' theorem, a theorem that details the relationship between conditional probabilities. Again, Read (2000) provides a simple and elegant explanation of this relationship.

Consider the probability of A being true, p(A). Now, consider the proposition, p(A|B) as the probability that A is true given that B is true. In other words, what we can say about A is related to what we know about B. There is an overall probability of A being true, and there is a conditional probability that A is true given what we know about B. For example, consider the "probability space" of A and B to be represented by the Venn Diagram in Fig. 7.7.

The area occupied by A is the p(A), and the area occupied by B is the p(B). Note that the two areas overlap: exactly 50 percent of the area of A is included within the area of B and 40 percent of the area of B is included within A. That overlap is the area where both A and B are true. That area of A that does not overlap the area covered by B is the area where A will be true even if B is false. The joint probability can be derived from the multiplication of probabilities:

$$p(\text{A and B}) = p(\text{A})p(\text{A given B})$$

In the Venn diagram, A occupies 20 percent of the probability space, so: p(A) = 0.2. This is the probability of A. Likewise, the probability space occupied by B is 25 percent, so: p(B) = 0.25. This is the probability of B. The region covered by both is p(A and B) = 0.1. Now, if we assume that A is true, then B will also be true half of the time:

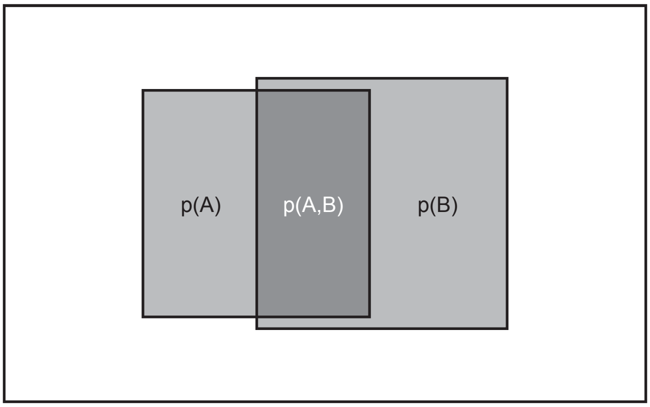

Figure 7.7. The probability of A + B is the joint probabilities of A and B. Redrawn from Read (2000), used with permission.

$$p(\text{B given A}) = 0.1/0.2 = 0.5 = p(\text{B and A})/p(\text{A}) = 0.5$$

This is the probability of B being true given A. If we assume that B is true, then 40 percent of the time A will also be true:

$$P(\text{A given B}) = 0.1/0.25 = 0.4 = p(\text{A and B})/p(\text{B})$$

This is the probability of A given B. Now, we can verify the multiplication law of probabilities:

$$p(\text{A and B}) = p(\text{A})p(\text{B given A}) = (0.2)(0.5) = 0.1$$

$$p(\text{A and B}) = p(\text{B})p(\text{A given B}) = (0.25)(0.4) = 0.1$$

Thus:

$$p(\text{A and B}) = p(\text{a})p(\text{B given A}) = p(\text{B})p(\text{A given B})$$

and

$$p(\text{B given A}) = p(\text{B})p(\text{A given B})/p(\text{A})$$

Now, consider one of these “B” to represent a model and “A” to represent data. We can cast the formula in likelihood terms:

$$p(\text{model given the data}) = p(\text{model})p(\text{data given the model})/p(\text{data})$$

and in more familiar notation

$$p(\text{model} \mid \text{data}) = p(\text{model})p(\text{data} \mid \text{model})/p(\text{data})$$

The denominator, p(data), is actually complex and is the sum of the probability of the hypothesis multiplied by the probability of the data, given the hypothesis,

$$P(\text{model} \mid \text{data}) = P(\text{model})p(\text{data} \mid \text{model}) / \Sigma_{\text{model}}\, p(\text{model})p(\text{data} \mid \text{model})$$

The denominator performs a desirable function: it normalizes the posterior probabilities so that they add to one. The problem is: the denominator is the sum of all models, and for a phylogenetic hypothesis that is only moderately large there are a great number of models—too large to be calculated. Thus, it is not just the tree topologies that can reach astronomical numbers but also the branch lengths. A complete account of all of the models for a particular tree topology would be all the possible branch lengths for that topology, which are a function of all of the rate parameters. Multiply that by the number of possible trees, and you can see why we need to approximate, rather than calculate, the answer.

Now, let us consider our original very simple example from the Bayesian perspective. Given our data, we can see intuitively that if the ML is reached with a model of mean = 5.0 mm and standard deviation = 1.0, the probability of a model that specified mean 4.9 mm, standard deviation 1.1, given the data, would be higher than the probability of a model with the mean = 6.0 mm and the standard deviation = 2.0. The question is: if we do not know the mean and only have the measurements, how would we reach the solution in a Bayesian manner? Consider the landscape showing our contours, Fig. 7.8. In a Bayesian analysis, the shapes and contours of this landscape are not known; we have to discover them.

If we performed all of the likelihood calculations at, say intervals of 0.1 mm × 0.1 mm, we would have one term in the equation. We also need a prior, the p(model), to complete the equation. This requirement for a prior partly leads to controversy because frequentists might claim that there is no justification to introduce prior expectations into the calculation. Classical phylogeneticists might also object to the statistical nature of this kind of inference. Still, even in classical phylogenetics, concepts akin to priors exist such as designated outgroups used to polarize characters.

If we were sampling the population a second time, the prior for the mean might be centered around 5.0 and the prior for the standard deviation might be a distribution centered around 1.0. If we have never sampled the population, perhaps we don't expect anything in particular. We might consider each fitted model to be equally probable. Since we have dissected the landscape into little squares, we can calculate the probability of each square and the entire lot of these calculations will sum to p =1.0. This creates a hill whose volume is 1 (Fig. 7.8). If we project the volume under each square, the sum of these volumes = 1, so the volume under each square is a measure of p(model|data) and the little square with the most volume happens to be the one at the top of the cone which includes mean = 5 mm, standard deviation = 1. The shape of this hill can tell us a great deal about the model as a whole and cross-sections can tell us a great deal about the model parameters.

The shape of a hill is a function of the credible intervals we might expect for any parameter. If there is only one peak (as in our case, cross-section Fig. 7.8a), then we can conclude that the problem is rather simple. If the surface has multiple peaks, then we can conclude that the fit of different models to the data is complex. If the cone forms a steep peak (cross-section Fig. 7.8b), then the difference in volume

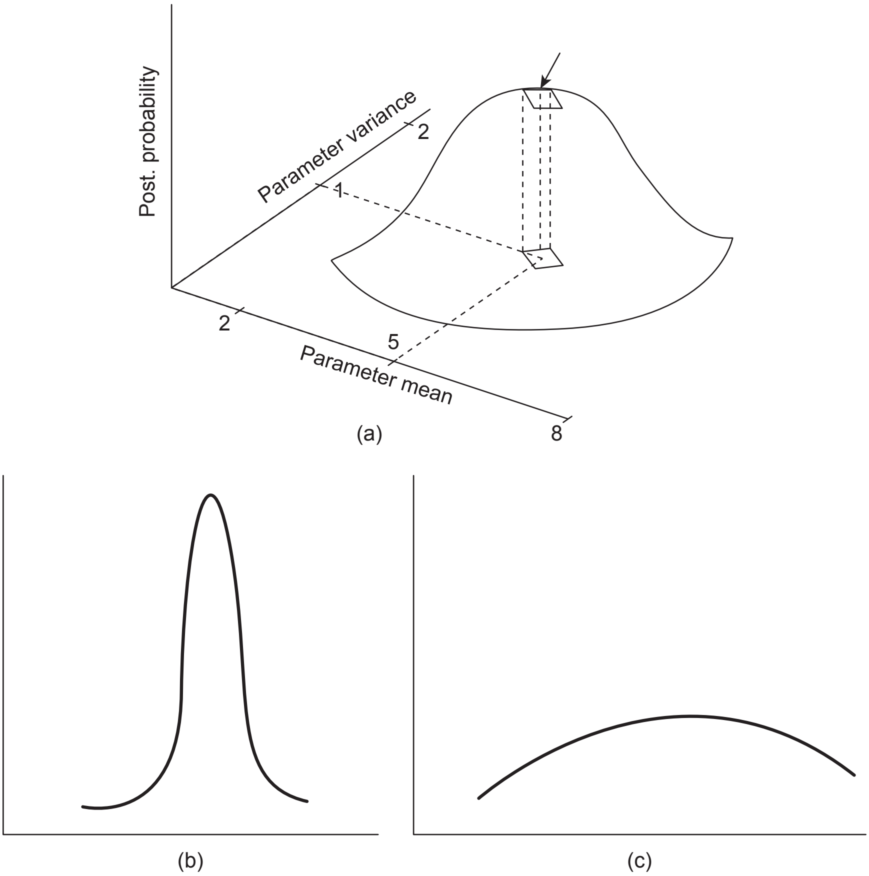

Figure 7.8. (a) Posterior probability space for our example of antenna length from Fig. 7.2. (b–c) Cross-sections of two probability spaces. In (b) there are very few fitted models at the top of the hill that have similar likelihood values. In (c) there are many fitted models at the top of the hill that have similar likelihood values.

among adjacent squares is large and we will have relatively few fitted models on the peak that are statistically equally probable. If the cone is in the form of a low hill (cross-section Fig. 7.8c), then the difference in volume among adjacent squares is very small and we have many models that might be equally probable. This result is very intuitive if we apply frequentist statistical reasoning. We know that confidence limits in frequentist statistics are a function of variation in the population. We may estimate a sample mean, and if the variance is low, we can expect to sample something similar the next time we sample the population for the same number of individuals. However, if variance is very high, then we might expect very different results the next time we sample, and our confidence would be low.

We can also isolate parts of the model, concentrating on only one model parameter and treating the other parts as nuisance parameters. For example, if we were only interested in the mean and its credible interval, we could treat the standard deviation as a nuisance parameter, and we could sum the probability along the axis that corresponds to the standard deviation. This would give us a one-dimensional curve showing the posterior probability density as a function of the mean. We can examine this curve to learn about what values for the mean are most probable in terms of posterior probability. We can do the same thing for the standard deviation.

In tree inference, we are interested in the tree topology that has the highest posterior probability. Any particular tree topology is associated with many other parameters, and it is theoretically possible to examine the posterior probabilities of each tree topology in a manner similar to examining all posterior probabilities of fitted mean values in our simple example. Some trees will be more probable, given the model, than others. And we can examine the posterior probabilities of other parameters, one at a time, given a particular tree topology, the other parameters, and the model, which is something of interest to evolutionary biologists interested in studying various evolutionary processes.

Now, we mentioned that summing all of the values in the denominator of Bayes' theorem was hard. In fact, it is an intractable problem in phylogenetic inference, due to the numbers of trees, branch lengths, and other parameters mentioned above. So, we cannot visualize the posterior probability surface or even visualize the complexity of the landscape to see if there are multiple optima. So, what to do?

As it turns out, this is not an insurmountable problem in phylogenetic inference because we can approximate using a variety of heuristic methods. The favored method used in Bayesian phylogenetics is to explore the possible space occupied by the models. If enough space can be visited, then perhaps we can find the hill and its peak without visualizing the entire landscape. This is somewhat analogous to exploring parsimony space using such routines as branch swapping. In parsimony we explore a landscape of tree lengths, and in Bayesian analyses we explore a landscape of posterior probabilities. In a parsimony analysis, we cannot see the entire landscape. We simply try to find the highest peak(s), and when we do, we accept the results as our best estimate. This is analogous to Bayesian analysis where we cannot see the entire landscape, but try to find the highest peak. Exploring model space is somewhat similar, except for the fact that we are adopting different criteria (tree length versus posterior probability).

There are several ways to explore the space. For example, Rannala and Yang (1996) used numerical integration of the posterior probabilities over all tree interior nodes, an approach that could be viewed as similar to an exhaustive search in parsimony, and one that suffers from the same defect: it is only useful for problems involving small numbers of taxa. The favored method is Markov Chain Monte Carlo (MCMC) integration.

Monte Carlo methods were first proposed by the mathematician Stanislaw Ulam (1909–1986: Weisstein, 1998). A simple description of how MCMC works is given by Lewis (2001a). In general, MCMC involves using computer-generated random numbers and a set of rules to simulate a walk through the space of trees and parameters. One begins by either randomly picking a model (random tree topology and other associated parameters) or by picking a particular model (one considered a priori probable, usually a particular tree topology and associated parameters). One then randomly picks a second model and compares it to the first. If the proposed model has a higher posterior probability density, then adopt it and pick another random likelihood model to test. But if the proposed model has a lower posterior, then it can still be picked with some probability (the probability is simply the ratio of the posterior for the proposed state to that of the current state). For example, if the second model is actually better than the first, then we pick the second 100 percent of the time. However, if the posterior probability density of the second tree is 75 percent of the current tree's posterior probability density, then pick the second tree 75 percent of the time by random draw, otherwise we stick with the original tree, pick a new tree model, and make a new comparison.

These methods are described by Metropolis et al. (1953) and Hastings (1970). If MCMC procedures are followed correctly, we will travel over the landscape and (given a long enough simulation) the amount of time the simulation spends in a particular set of models will approximate the posterior probability of that set. If we save the results as a function of the frequency with which we sample various models, the procedure can sample the shape of the hill and allow us to find the approximate location of the peak. This procedure is sampling the *posterior probability density* of the models given the data. It works rather like an n-dimensional histogram, the more probable models are visited more often than the less probable models, building a probability density space where the best model(s) are interpreted as more probable because MCMC visits them more often ("finds them") than the less probable models.

MCMC searches run forever unless stopped by the investigator. In phylogenetics, the end result we seek is a stable topology or set of topologies, so all other parameters (e.g., those determining branch lengths) are treated as nuisance parameters. (Of course, if the objective is not strictly a systematic objective, then other parameters may be of interest.) If we run MCMC long enough, we will be able to accurately estimate the posterior probability of any tree. If the peak is steep and pointed, then the area of the peak is very small. If we could visualize its volume relative to the entire landscape, the volume would be highest. In such a situation, we would expect the peak to be populated by a relatively small set of trees and associated parameters (Fig. 7.8b). If the peak is low and flat, it will be populated by a large number of tree topologies and associated parameters (Fig. 7.8c).

Consider a simple thought experiment. Pretend that you have traveled around and over the surface for a million iterations. That means you have visited 1×10^6 models, with replacement. One model may be visited many times, another only a few times. If all of the states you have visited corresponded to the same tree topology, then 100 percent of the time you would have encountered the same clades. Consider another case in which you visit 15 tree topologies. In this case, it is possible that no single tree dominates the posterior probability surface, but you can still summarize the proportion of steps in which you were in a state that corresponds to each tree. This will allow you to estimate the posterior probability of each tree. You can also calculate the proportion of steps in which you visited a tree that had a particular clade. This proportion is interpreted as the posterior probability that this clade is present in the true tree.

The usual procedure is as follows.

1. Select a model that contains a reasonable set of parameters for change in the same manner as selecting a likelihood model for ML. In the best of all possible worlds, this would be the true model of character evolution, but in the real world, one hopes that the analysis is robust to the use of a simpler model.

2. Run an MCMC analysis. This consists of one or more Markov chains (four is common), each sampling the probability landscape. The point is to sample the posterior probability space as thoroughly as possible and find the area(s) of highest probability density. There may be multiple peaks of different heights. So the typical analysis may consist of multiple MCMC chains and criteria for switching from one chain to another, if it looks more promising.
3. The first MCMC results are likely to be uninteresting because the analysis is just beginning to sample the probability space and is heavily influenced by the starting point for the MCMC (and this starting point is often arbitrary). Frequently the initial portion of the MCMC run is discarded as a "burn-in." Convergence of the chain to accurate approximation of its stationary distribution is difficult to detect. Multiple independent runs are usually required. Discrepancies between different simulations with respect to the estimates of the posterior probability surface indicate that the simulations (or at least some of the runs) were terminated prematurely.
4. If tree topology is the parameter and all other parameters are considered nuisance parameters, the goal is to determine the relative frequency of the appearance of particular clades in the probability space sampled. *The frequency at which you sample any clade provides an estimate of its posterior probability.* It is possible to collect all of the samples, but this would take a great deal of computer space because we would have to save upward of 1 to 3 million tree topologies, one for each visit to the peak by MCMC. More typically, the investigator saves some subsample, typically one tree per thousand sampled. Consensus analysis can then be performed on the collection of trees. The usual method is to take a majority rule consensus tree, which will reveal the percentage of times a particular clade appears in the subsample and this is interpreted as the posterior probability of the monophyly of the clade given the data.
5. To understand the results, it is helpful to consider what might be on the peak. Consider first a very simple problem: a set of data having little homoplasy. If we ran a parsimony analysis, we would quickly find a solution and, if the data are really clean, perhaps only a single tree that is optimal. In this case, we would also expect the same result using likelihood methods, or classic Hennigian argumentation with a priori character polarization. In such a case, we can imagine that the peak is very pointed and at the top are a rather large number of models that do not vary in tree topology, but might be different in the nuisance parameters of branch length, a function of the Q-matrix. (Indeed, there are, potentially, an infinite number of such models because branch length variation is continuous.) In this case, you would expect the algorithm to sample the same clades, over and over again, discovering alternative topologies very infrequently. Now consider a very messy problem: a great number of taxa and many potential homoplasies. A parsimony analysis might result in many parsimonious trees (or many trees that differ only slightly in length), perhaps collapsing into one big polytomy if a consensus analysis was performed. Our confidence in clades that appear only in some subset of the most parsimonious trees is not very high; bootstrap values are low for most clades. In this case, you would expect the peak to be broad and populated with many topologies

> plus the complement of diversity in nuisance parameters. You would expect the algorithm to sample any particular clade at a low frequency or perhaps not at all.

There are several computer packages available for performing Bayesian phylogenetics. The most commonly used package seems to be Mr. Bayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). An excellent introduction to Bayesian literature is presented on the Mr. Bayes WWW site: mrbayes.csit.fsu.edu.

INTERPRETING MODELS IN A PHYLOGENETIC CONTEXT

It is important to understand that likelihood methods may be used as tools in both systematic research and in broader evolutionary research. Strictly speaking to systematic research, parameters such as branch length are of concern only when branch lengths interfere with reconstructing common ancestry relationships. This can be a concern when taxon sampling or character evolution results in "long branch attraction" problems (see, for example, Anderson and Swofford, 2004). Otherwise, phylogenetic systematics, as a discipline, takes no account of branch lengths because branch lengths that are not artifacts of "long branch attraction" are not taken into account in determining genealogical relationships or in forming phylogenetic classifications. Trying to account for such things in classifications is a research program for evolutionary taxonomists, not phylogeneticists. If, however, the thrust of the research is in other evolutionary directions, then we may wish to have a tree before the fact and study the evolutionary behavior of characters on that tree. Such research programs are not primarily systematic, but depend on the fruits of systematic analysis.

There are two potential advantages of using explicit models in a likelihood-based framework. First, likelihood-based techniques are parametric techniques, and parametric techniques are usually more powerful than nonparametric techniques. *Power*, in this sense means statistical power, the ability of the statistical test to correctly reject a false null hypothesis. Second, if the analysis uses the correct model of evolution, then ML will have the characteristic of statistical consistency. *Statistical consistency* obtains when the analysis converges on the correct solution as more and more data of the same kind are applied to the problem. Early proponents of likelihood touted statistical consistency as a clear advantage of likelihood over parsimony (e.g., Felsenstein, 1978). Parsimony adherents pointed out that there was no guarantee that ML methods would produce statistically consistent results unless the model was true, and that because no one claimed that their models were true, the claim is unjustified (Farris, 1999). ML adherents replied that ML was robust to deviations in model truth-value, which seems to be reasonable under certain conditions, but not in others, especially when evolutionary rates are heterogeneous across or within sites and when these parameters are not included in the model (e.g., Gaut and Lewis, 1995). Sober (2008) states that scientists use models they know to be "untrue" all of the time and that models should be judged on their power to predict new data such that fitted models can be tested. Sober (2008:351–352) suggests the controversy between parsimony and likelihood is not likely to be solved by simply process models where parsimony and likelihood agree (e.g., Felsenstein, 1981a; Penny et al., 1994; Tuffey and Steel, 1997) or disagree because parsimony advocates will claim that the models are wrong or unknowable and likelihood advocates will claim that the models are at least good enough to make accurate predictions.

Given this acrimony, we were curious to see exactly what might emerge when we performed both a parsimony analysis and a Bayesian analysis on a large morphological data set and then used the ability of Mesquite (Maddison and Maddison, 2009) to plot the distribution and probabilities of likelihoods of synapomorphies on the tree topology generated from the Bayesian analysis and compare it to the results of the parsimony analysis. We could not find argumentation. The question is: does only parsimony, as an algorithm, fulfill the basic principle of Hennig (1966) that one should not assume convergence in the absence of evidence? An additional question: does a likelihood approach result in monophyletic groups confirmed by synapomorphies? The auxiliary principle and grouping by synapomorphy are thought by us to underlie all phylogenetics. But what if parsimony and classical Hennigian argumentation are not the only way to fulfill the paradigm? Or to put it another way, is there more than one way to be a phylogeneticist in the Hennigian tradition?

With the help of Matthew Davis (then a graduate student at the University of Kansas), we analyzed the Gauthier et al. (1988) matrix of amniote relationships (fossil and living) using both parsimony (PAUP*, TBR, 100RAS, equal weighting) and a Bayesian analysis with the simple Mk morphology model of Lewis (2001b). We did not perform a likelihood analysis as there is no way at present to examine synapomorphies at nodes using likelihood at this stage in the development of algorithms. Ancestral states reconstruction, as implemented in Mesquite (Maddison and Maddison, 2009), was used to study the distributions of characters assigned to ancestral nodes. Because of missing data, we did not study the state reconstructions of the soft anatomical characters, but they were included in the analyses. The single tree topologies of both analyses were identical. Of the 207 hard anatomical characters studied, 198 unambiguous synapomorphies were mapped to the same node in both the parsimony and Bayesian tree topologies. The only difference was that the Bayesian probabilities were on the order of 95 percent to 99 percent while those of parsimony scored CI values of 1.0. Characters that were ambiguous in the parsimony analysis were ambiguous in the Bayesian analysis (9 of 9) and at the same nodes, and there were no character columns that were totally different in their interpretation (although 7 instances of states were different, e.g., probabilities of 75 percent in Bayesian versus CIs of 1.0 in parsimony). Of course, this is only one analysis and one Bayesian model, but these results seems to show that at least for one matrix the Bayesian analysis is attempting to maximize homology and minimize homoplasy, and it is circumscribing monophyletic group with synapomorphies—two of the goals of the Hennigian Paradigm. We conclude that parsimony may not be the only way to achieve classical Hennigian objectives.

CHAPTER SUMMARY

- Likelihood asks the question: what is the probability of observing these data given a specified model (which includes a tree topology)?
- In phylogenetic inference, the model includes a tree topology and some number of other parameters including branch length between nodes.

- Likelihood models take into account the probability that characters will change over time between nodes.
- The simplest model is picked from a variety of models available.
- Bayesian analysis asks the question: what is the probability of the model (including the tree topology) given the data?
- Bayesian analysis also includes both the tree and other parameters such as branch length.
- Bayesian analysis uses the formalities of likelihood calculations and explores posterior probability space using MCMC to find the model/tree topology with the highest posterior probability.
- For phylogenetic research, branch lengths are of concern only in so far as they affect our ability to estimate genealogical relationships, so the important part of the model in both kinds of analyses is the tree topology.