

Phylogenetic Inference with Distance, Maximum Likelihood, and Bayesian Methods

The previous chapter presented Hennigian inference and parsimony as a means to illustrate the principles underlying phylogenetic analysis in general. However, Hennigian analysis is not in current use, and parsimony, while widely used, is no longer the most common method for phylogenetic inference. In this chapter, we will introduce three other approaches to reconstructing trees. The first, based on evolutionary distances, is frequently used in situations where speed of analysis rather than phylogenetic precision is important. The other two approaches, maximum likelihood and Bayesian analysis, are sophisticated methods that use mathematical models of character evolution to achieve more precise estimates of phylogenetic history. Because of their statistical power, maximum likelihood and Bayesian analysis are the major methods used in phylogenetic research, at least for DNA sequence data. Before describing these three methods, it will be useful to introduce mathematical models of evolution because these provide a basis for a majority of the methods described in this chapter.

INTRODUCTION TO MODELING MOLECULAR EVOLUTION

In Chapters 4 and 7, we introduced the outlines of a model for the evolution of a discrete character along the branches of a phylogeny. We stipulated that characters evolve independently and that they may switch among a finite number of character states. We also specified that traits evolve along the branches of a tree,

and that these branches have a defined (but maybe unknown) duration. These factors provide a sufficient basis for visualizing trait evolution and for seeing why a method such as parsimony might work. However, to really understand trait evolution and to apply model-based approaches such as maximum likelihood and Bayesian inference, it is necessary to add more mathematical details to this verbal model.

The core of any mathematical model of character evolution is a *substitution model*, which specifies the way in which characters are permitted to evolve between states as well as the relative rates of different kinds of evolutionary change. All models in widespread use in phylogenetics are continuous-time *Markov models*; that is, they describe a process in which the probability of an event happening in some time window is dependent only on the state at that time and independent of how it came to be in that state. Coin tossing is usually modeled as a Markov process. In such a model there is some *fixed* rate at which heads show up (probably 0.5 per toss). This means that the probability of heads is unaffected by the number of heads that were obtained in previous throws.

Mathematical models have now been developed for many different types of traits. We will focus on models of DNA sequence evolution. These are easy because there are only four possible states corresponding to the four bases, A, C, G, and T. While indels occur, these are not included in the basic models of sequence evolution because they add too much complexity. Instead, gap characters are treated as missing data, representing the fact that we are uncertain whether nucleotide positions that are absent should be scored as an A, C, G, or T.

We will also restrict our discussion to substitution models that are *time reversible*. This means that the probability of a change between two states is equal in the forward and reverse directions. For example, the number of changes from A to T is assumed to be equal to the number of changes from T to A. While time reversibility is not strictly required, it provides a reasonable simplification in cases where we think that base composition (i.e., the relative frequency of the four bases, A, C, G, and T) has not changed systematically over time.

We will assume that at any one moment in time a particular position in a DNA sequence is occupied by one of the four bases and that every so often a mutation occurs. The best way to visualize this mutation is as a two-step process: the old base is removed and then a base is drawn at random from a pool of possible bases and is inserted in the position of the removed base. If the inserted base were the same as the deleted one, no change would be visible. Only if one of the three other bases were drawn from the pool would a change be visible.

We will assume that both steps, and thus the entire mutational process, occur in zero time.

To get a feel for this process, we will begin with an analogy. Imagine a card sitting face-up on the table in front of you next to a deck of cards. We will pretend that these cards do not have numbers, but are characterized just by their suit. Now imagine an invisible card-changing fairy who occasionally removes the card on the table, replaces it in the deck, and then draws a card at random from the deck (a deck with an equal number of cards of each suit) to put back onto the table. Being a fairy, this process happens faster than the human eye can track. If the new card is the same suit as the old card, there would be no visible change, but otherwise the card would change suit. You can imagine watching a card on the table and seeing it occasionally changing to a new suit. Let us now develop a mathematical description of the way the card changes.

Suppose that although the fairy changes cards at a specific rate, being flighty, he does so in an unpredictable manner, analogous to the way that radioactive atoms decay. We will focus on how frequently the fairy changes the identity of a card's suit. We will ignore events where a card is replaced by another card of the same suit (although the underlying math does deal with these "hidden" events). Let us call the rate at which suits change the substitution rate, μ .

Suppose the fairy changes cards at a rate of 0.6 substitutions per minute. This would predict that if we watched a single card for an hour, we should see an average of 36 changes (0.6×60 minutes). However, the actual number would vary hour to hour and, even more so, minute to minute. Sometimes the fairy would change the card several times in quick succession, but other times he would wait a long time between changes. However, while the waiting time between successive changes will be variable, it will average 1 minute, 40 seconds ($= 1/0.6$ minutes). (For the mathematically inclined, the waiting time is exponentially distributed with a mean of $1/\mu$ minutes.)

If you watch the card for a long time, you will see it visiting all four suits. A useful way to keep track of the pattern of changes is with a substitution matrix, such as that shown in Figure 8.1. This shows the expected number of substitutions of each type seen in a certain amount of time. Of the 16 possible substitutions, only the 12 off-diagonal events result in a visible change in the card. Because each of the 12 possible changes should occur equally frequently, and because the overall rate of visible change is μ , the rate at which each of the 12 changes occurs is $\mu/12$. For example, if the overall rate of change were 0.6 changes/minute, then the rate at which any of the specific kinds of change

		To:			
		♠	♦	♥	♣
From:	♠	—	$1/12 \mu t$	$1/12 \mu t$	$1/12 \mu t$
	♦	$1/12 \mu t$	—	$1/12 \mu t$	$1/12 \mu t$
	♥	$1/12 \mu t$	$1/12 \mu t$	—	$1/12 \mu t$
	♣	$1/12 \mu t$	$1/12 \mu t$	$1/12 \mu t$	—

FIGURE 8.1 Expected number of changes by the card-changing fairy in t minutes. The overall rate of card evolution is μ substitutions per minute.

		To:			
		♠	♦	♥	♣
From:	♠	$-\mu$	$\mu/3$	$\mu/3$	$\mu/3$
	♦	$\mu/3$	$-\mu$	$\mu/3$	$\mu/3$
	♥	$\mu/3$	$\mu/3$	$-\mu$	$\mu/3$
	♣	$\mu/3$	$\mu/3$	$\mu/3$	$-\mu$

FIGURE 8.2 Instantaneous rates of substitution by the card-changing fairy. The overall rate of card evolution is μ substitutions per minute.

occurs should be $0.6/12 = 0.05$ changes per minute. Thus, if we watched a card for 1000 minutes, we would expect to see about 50 of each kind of substitution.

The substitution matrix reports the expected number of each kind of change as a function of the substitution rate, μ , and time, t . In order to make more useful predictions, we need to express this model of card “evolution” in terms of *instantaneous rates* of change. These are summarized in Figure 8.2. The entries in the matrix report the rate at which a card starting at the suit shown to the left will switch to each of the other suits. Since the rate of change is μ , and there are three possible alternative suits at equal frequency in the deck, the rate of change to each is $\mu/3$. The “rates” of staying in the present base are set to $-\mu$, which means that the sum of each row is zero. One way to think about this is that, because every change from the starting base results in a suit within the row, the net rate of leaving a row must be zero.

Before developing the fairy metaphor further, let us clarify the link to the goals of phylogenetic inference. Phylogenetic inference is essentially an attempt

to determine how long ago a pair of taxa last shared common ancestry. Rather than counting time in units of years, which can be done but adds complications (Chapter 11), we will focus on how far apart two taxa are from each other in substitutional units. That is to say, we attempt to determine their *evolutionary distance*, which for DNA sequence data is the average number of substitutions that have occurred at each nucleotide position. Returning now to the fairy metaphor, we shall see why it is necessary to analyze whole sequences, comprising multiple aligned nucleotide positions rather than just one position at a time.

Suppose we put a spade card on a table and leave the room for 10 minutes. Based on the suit of the card when we return, can we say anything about the number of changes that happened while we were away (the number of changes being analogous to evolutionary distance)? To answer this we need to know how the probability of observing each of the four suits changes as a function of the substitution rate, μ , and time, t . What we want is a *substitution probability matrix*, which shows the probability that a card starting at a particular suit would be found to be in each of the four possible suits t time units later.

A substitution probability matrix is not easy to figure out by hand because there are an infinite number of histories that start with a spade and end with, say, a heart. There could have been one change directly from a spade to a heart. But there could have been, for example, a change from spades to clubs, then a change from clubs back to spades, then a change from spades to diamonds, and finally a change from diamonds to hearts. Thankfully, with the help of calculus and matrix algebra, substitution probability matrices can be derived from instantaneous rate matrices. The substitution probability matrix for the fairy model is given in Figure 8.3. The base of natural logarithms, e , emerges when we mathematically integrate over the alternative possible histories.

		To:			
		♠	♦	♥	♣
From:	♠	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♦	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♥	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♣	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.3 Substitution probability matrix under a simple model of card “evolution” with equal frequencies of the four suits. The mutation rate, in card changes per minute, is denoted μ . The number of minutes over which evolution is allowed to happen is denoted t .

Before moving on, it is worth looking briefly at the matrix and trying to develop a feel for what it communicates. What is the probability of starting as a spade and ending as a spade some very short period of time later? If μt is small, then $e^{-\mu t}$ is close to 1.0 (remember, any number taken to the power of 0 has a value of 1). In that case, the probability of still being a spade is close to $\frac{1}{4} + \frac{3}{4}$, or 1.0, and the probability of being in any other state is almost zero. This makes sense. If no time has passed, you must still be in the state you started in.

Now consider the other extreme, when μt is very large. In that case $e^{-\mu t}$ is very close to zero, meaning that the probability of being in any of the four states is $\frac{1}{4}$, or 0.25. This too makes sense. If you wait an infinite amount of time, the ending suit is no longer constrained by the starting suit. In that case, with all the suits occurring at equal frequency in the deck, there is a 25% chance that the card now showing is of each of the four suits. For any time between 0 and ∞ , the probability of the card still being a spade is between 0.25 and 1.0, while the probability of the card being another suit is between 0 and 0.75. Table 8.1 lists some probabilities for different values of μt .

Now we can return to the original question: If we leave the card for 10 minutes, can we estimate the number of changes of suit that occurred while we were out of the room? If we knew the substitution rate, μ , then we would *know* the evolutionary distance, which is defined as μt , without even looking at the card when we return. If we did not know μ , then we would have a problem. The card is either the same or a different suit, but in either case there could have been few

TABLE 8.1 The probability of a card starting as a spade and being a spade or another suit after an average of μt substitutions have occurred (Probability of not being a spade = 1 – Probability of being a spade)

μt	Prob[♠]	Prob[not ♠]
0.01	0.990	0.010
0.05	0.952	0.048
0.1	0.906	0.094
0.5	0.635	0.365
1	0.448	0.552
5	0.251	0.749
10	0.250	0.750

(0 or 1) or infinitely many changes while we were out of the room, depending upon the value of μ . We cannot estimate μ by seeing if one card has changed state between two moments of observation.

The solution is to leave behind not one card but a whole line of cards. By looking at the proportion of the cards that had changed suit in 10 minutes, we can obtain information about μ , and this, in turn, allows us to estimate the number of changes of suit that occurred when we were not looking. Let us walk through this numerically.

Suppose we had laid out 100 cards in a row before leaving the room and noted their suit. If we came back and 60 cards had the same suit as before we left and 40 had changed, how many changes on average did each card experience? You might be tempted to say 0.4 changes per card (40/100), but this ignores all the substitutions that were then “covered up” by further changes. Using the substitution probability matrix, we can take account of the extra changes.

Since the proportion of cards that is unchanged is 0.6, our best estimate is that the probability of not changing suit is also 0.6. The substitution probability matrix shows us that we just need to find the value of μt such that $\frac{1}{4} + \frac{3}{4}e^{-4/3\mu t} = 0.6$. Some simple algebra solves the equation and finds that the evolutionary distance, μt , is 0.572. In this case, because we know that t is 10 minutes, we can also calculate the value of μ (0.57/10 = 0.057 substitutions/minute). Figure 8.4 shows how the frequency of cards that *have* changed suit

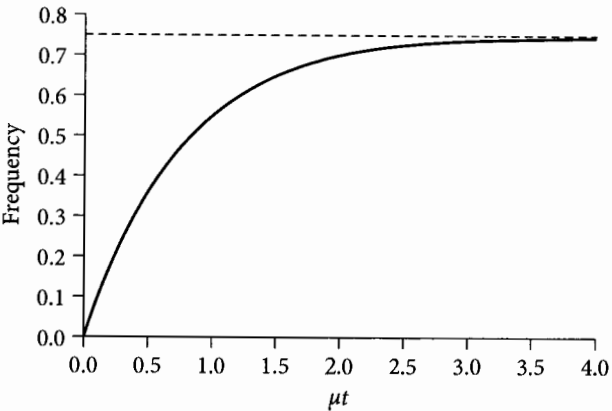


FIGURE 8.4 Expected frequency of cards that have a different suit from the ancestor as a function of μt .

changes as a function of μt . You will see that a frequency of 0.4 corresponds to $\mu t = 0.57$. This shows that if we are willing to assume that all the cards are acted upon identically by the fairy, then by looking at the proportion of suits that change between two moments of observation, we can estimate the number of hidden substitutions that happened and thereby calculate the card analog of evolutionary distance: the average number of changes of suit per card.

The card-changing fairy metaphor can now be related back to DNA sequence evolution. The four suits correspond to the four bases, A, C, G, and T. An individual card corresponds to a single position in a DNA sequence, and a line of cards corresponds to a sequence of DNA. The simple model illustrated with the fairy was originally developed by Jukes and Cantor and is usually called the Jukes-Cantor or JC model of molecular evolution (Jukes and Cantor 1969). This model assumes that (a) all four bases occur at equal frequency, (b) each kind of substitution (A to C, A to T, etc.) occurs at an equal rate, and (c) the rate of substitution is the same for all nucleotide positions in the sequence being studied. Under these assumptions, the substitution probability matrix for nucleotide positions and bases resembles that for cards and suits, as shown in Figure 8.5.

Let us think about what would happen if you let an ancestral sequence evolve while you kept track of the proportion of positions at which the sequence differs from its ancestor. At the beginning, the sequence will be identical to the ancestor, but differences will gradually accumulate. The initial rate of increase will be equal to μt because each change adds to the difference between the ancestor and descendant. However, as time continues, more and more changes will fail to generate additional differences. For example, if the ancestor had a G, which subsequently changed to a C, a later change to a T at this position would not be

		To:			
		A	C	G	T
From:	A	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	C	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	G	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	T	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.5 Substitution probability matrix under the JC model of DNA sequence evolution. The mutation rate, in substitutions per unit time, is denoted μ . The time interval over which evolution is allowed to happen is denoted t .

a new difference. Furthermore, some changes can actually reduce the difference by bringing the descendant back to the state found in the ancestor. Thus, over time the rate of increase of distance will slow down until eventually the distance is “stuck” at 0.75 (see Figure 8.4).

MORE REALISTIC MODELS OF
MOLECULAR EVOLUTION

While the JC model provides an accessible starting point for thinking about models of molecular evolution, all three of its core assumptions are violated by real DNA sequences. The four bases are usually not present at equal frequencies, some kinds of substitutions occur at different rates than others, and some positions in a DNA sequence have a higher rate of evolution than others. Over the last four decades, more sophisticated models have been developed that more accurately reflect these and other violations of the JC model.

The first extension of the JC model is to allow bases to occur at different frequencies. This model is usually called F81, because of when and by whom it was proposed (Felsenstein 1981a). It is analogous to assuming that the deck from which the fairy draws the card does not necessarily have an equal proportion of the four suits. There are several possible reasons why base composition is expected to be uneven. For example, selection can favor a higher frequency of guanine and cytosine in some RNA molecules (e.g., ribosomal RNAs) because these two bases form three rather than two hydrogen bonds, resulting in a more stable secondary structure. Whatever the cause, uneven base composition affects both the probability that a site will start in a certain base and the probability that it will be replaced by each of the other bases. Let us go through this more fully to illustrate the ways in which a basic model can be extended to include more biological realism.

Let us use π to represent the frequency of each base in the “pool” (analogous to the deck of cards) from which bases are drawn. Since there are only four bases, the base frequencies ($\pi_A, \pi_C, \pi_G, \pi_T$) add up to 1.0. Assuming that a sequence starts in equilibrium (where the frequency of bases in the sequence matches the frequency in the pool), the probability of a position being in state G at time 0 is equal to π_G . After a mutation has happened at a site, the probability that the new base will be a G is, again, π_G . This means that the expected number of mutations of a given type is μt times the product of the frequencies of the starting and ending bases. The matrix in Figure 8.6 shows the expected

		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	—	$\pi_A\pi_C\mu t$	$\pi_A\pi_G\mu t$	$\pi_A\pi_T\mu t$
	C (freq = π_C)	$\pi_C\pi_A\mu t$	—	$\pi_C\pi_G\mu t$	$\pi_C\pi_T\mu t$
	G (freq = π_G)	$\pi_G\pi_A\mu t$	$\pi_G\pi_C\mu t$	—	$\pi_G\pi_T\mu t$
	T (freq = π_T)	$\pi_T\pi_A\mu t$	$\pi_T\pi_C\mu t$	$\pi_T\pi_G\mu t$	—

FIGURE 8.6 Expected numbers of each type of substitution under the F81 model of DNA sequence evolution. The frequency of each base (A, C, G, and T) is indicated with the subscripted notation π , where $(\pi_A + \pi_C + \pi_G + \pi_T = 1)$.

		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	$-m(\pi_C + \pi_G + \pi_T)$	$\pi_C m$	$\pi_G m$	$\pi_T m$
	C (freq = π_C)	$\pi_A m$	$-m(\pi_A + \pi_G + \pi_T)$	$\pi_G m$	$\pi_T m$
	G (freq = π_G)	$\pi_A m$	$\pi_C m$	$-m(\pi_A + \pi_C + \pi_T)$	$\pi_T m$
	T (freq = π_T)	$\pi_A m$	$\pi_C m$	$\pi_G m$	$-m(\pi_A + \pi_C + \pi_G)$

FIGURE 8.7 The instantaneous rate matrix under the F81 model of DNA sequence evolution. Base frequency notation is the same as Figure 8.6. The effective mutation rate, after correcting for base compositional inequality (see text), is denoted m .

frequency of all 12 kinds of change. If you have studied genetics, you will note a resemblance here to the Punnett square method of predicting genotype frequencies. You will also observe that the expected number of A to G changes, $\pi_A\pi_G\mu t$, is the same as the reverse, $\pi_G\pi_A\mu t$. This means that, at equilibrium, the base frequency will tend to remain unchanged. It also means that the F81 model (like the Jukes-Cantor model) is time reversible: evolution looks the same whether it runs forward or backward in time.

The F81 instantaneous rate matrix (Figure 8.7) is derived similarly to the Jukes-Cantor model. Because the frequency of the starting base does not matter (the matrix shows the rate of a substitution conditioned on the identity of the starting base), the rates are influenced only by the frequency of the ending base. The diagonals are again such that the rows add up to zero. To be strict, it should be noted that the m included in this matrix is a modified version of μ used in the JC model. This modification is needed to account for the effect

of base frequencies on the instantaneous rate of substitution. Rare bases will tend to persist for less time (manifested as a higher rate of substitution) than common bases because common bases will often be substituted by themselves, resulting in no actual change. While it is not necessary to know the derivation of m to understand the principles, we provide the formula for completeness: $m = \mu(1 - \pi_A^2 - \pi_C^2 - \pi_G^2 - \pi_T^2)$.

The substitution probability matrix for the F81 model is shown in Figure 8.8. You will see that the probability of a substitution in each direction, for example, C to G versus G to C, can be different because the probabilities depend on the frequencies of the bases. This may seem to be at odds with Figure 8.6, which shows time reversibility (the expected number of substitutions in each direction is the same). The apparent discrepancy is explained by the fact that when a position is occupied by a rare base, it will tend, on average, to quickly switch to another state, but when a common base is present, it will tend to persist longer. This difference in the waiting time results from the fact that when a mutation event happens, it is relatively probable that the common base will be replaced by itself, resulting in no actual substitution.

For the mathematically inclined, you may notice that if $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$, then $m = 3\mu/4$ or $\mu = 4m/3$. If you plug this into the diagonal entries, they become $1/4 + 3/4e^{-4/3\mu t}$ and the off-diagonal values become $1/4 - 1/4e^{-4/3\mu t}$. These are identical to the JC model, showing that JC is a special case of F81 in which the four bases are at equal frequency.

The next level of complexity that molecular models consider involves allowing for different relative rates for different kinds of substitution. For example, suppose that the fairy had a propensity to replace a card by a suit of the same

		To:			
		A	C	G	T
From:	A	$\pi_A + (1 - \pi_A)e^{-mt}$	$\pi_C(1 - e^{-mt})$	$\pi_G(1 - e^{-mt})$	$\pi_T(1 - e^{-mt})$
	C	$\pi_A(1 - e^{-mt})$	$\pi_C + (1 - \pi_C)e^{-mt}$	$\pi_G(1 - e^{-mt})$	$\pi_T(1 - e^{-mt})$
	G	$\pi_A(1 - e^{-mt})$	$\pi_C(1 - e^{-mt})$	$\pi_G + (1 - \pi_G)e^{-mt}$	$\pi_T(1 - e^{-mt})$
	T	$\pi_A(1 - e^{-mt})$	$\pi_C(1 - e^{-mt})$	$\pi_G(1 - e^{-mt})$	$\pi_T + (1 - \pi_T)e^{-mt}$

FIGURE 8.8 Substitution probability matrix under the F81 model of DNA sequence evolution. Base frequency notation is the same as Figure 8.6. The effective mutation rate, after correcting for base compositional inequality (see text), is denoted m . The time interval over which evolution is allowed to happen is denoted t .

color (even beyond any bias that might arise based on the frequency of the four suits). This would mean that red-to-red and black-to-black substitutions would tend to happen at a higher rate than red-to-black or black-to-red substitutions. This turns out to be very similar to what happens biochemically to DNA during the mutational process because of the structural differences between purine (A and G) and pyrimidine (C and T) bases. As a result, transitions (purine-to-purine or pyrimidine-to-pyrimidine substitutions) tend to happen at a higher rate than transversions (purine-to-pyrimidine or pyrimidine-to-purine substitutions). This phenomenon is termed *transition:transversion bias*.

The method for accommodating this inequality (or any other) in substitution rate is to include a parameter in the model that adjusts the rate of one class of change relative to the other(s). In the case of the HKY (Hasegawa, Kishino, and Yano 1985) model, which includes both unequal base frequencies and transition:transversion bias, the relative rate parameter, denoted κ , is added to the rate matrix as a multiplier to the transitions (Figure 8.9): the higher the value of κ , the higher the rate of transitions relative to transversions.

This idea can be extended by allowing the two different kinds of transitions and/or the four different kinds of transversions to have different rate modifiers. This is achieved by adding parameters to the rate matrix that indicate the rate of certain substitutions relative to others. The most extreme case that is commonly used is the general time-reversible model, or GTR. This adds rate multipliers to five of the six rates of change (the six off-diagonal elements in the matrix). The sixth change does not need a multiplier because it is implicitly set to a rate of 1.0. Being time reversible, the changes in both directions (e.g., A to G and G to A) use the same rate multiplier.

		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	$-m(\pi_C + \kappa\pi_G + \pi_T)$	$\pi_C m$	$\pi_G \kappa m$	$\pi_T m$
	C (freq = π_C)	$\pi_A m$	$-m(\pi_A + \pi_G + \kappa\pi_T)$	$\pi_G m$	$\pi_T \kappa m$
	G (freq = π_G)	$\pi_A \kappa m$	$\pi_C m$	$-m(\kappa\pi_A + \pi_C + \pi_T)$	$\pi_T m$
	T (freq = π_T)	$\pi_A m$	$\pi_C \kappa m$	$\pi_G m$	$-m(\pi_A + \kappa\pi_C + \pi_G)$

FIGURE 8.9 The instantaneous rate matrix under the HKY model of DNA sequence evolution. Notation is the same as Figure 8.7 except for the addition of a rate multiplier, κ , which indicates how many times faster transitions occur than transversions.

The GTR model includes more free parameters than any of the other models we have described. Nonetheless, GTR does not require that there be an actual deviation from equal base frequencies, nor that the five substitution types be different from one another. The JC, F81, and HKY are all special cases of the GTR model or, put another way, they are nested within the GTR model. The HKY model is derived from GTR when the four transversions are equal and the two transitions are equal. F81 is derived when, in addition, the rates of transversions equal the rates of transitions. And, finally, JC is derived when, in addition, the base frequencies are equal. The connections among these models are shown in Figure 8.10.

The final major assumption of the models discussed so far is that all sites in a DNA sequence evolve at the same rate. However, we commonly expect

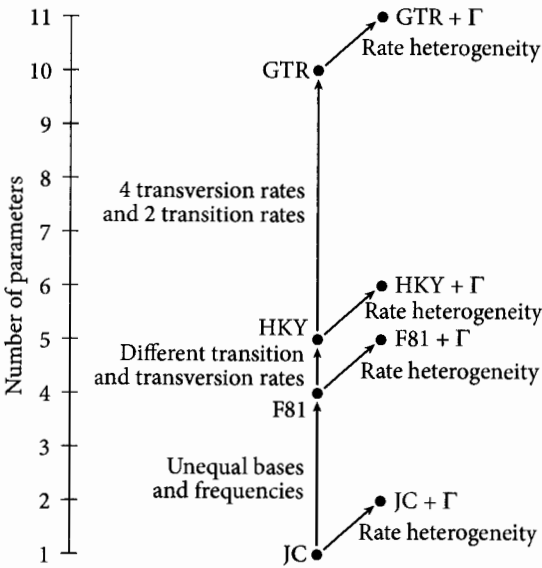


FIGURE 8.10 Depiction of the relationship between some commonly used models of evolution. Any two models that are connected by arrows that proceed in the same direction are nested: the simpler model (closer to the bottom of the chart) contains a subset of the free parameters of the more complex model. The axis on the left shows the number of free rate parameters in the model. The figure assumes that site-to-site rate heterogeneity is modeled using a discrete approximation to a gamma (Γ) distribution, which adds one free parameter to the model.

variation in rates among sites. This is expected given that selection will tend to slow down the rate of evolution of positions that critically affect gene function, which results in a higher rate of extinction of alleles that have variants at those sites. But how can we allow for rate variation while still being able to use the frequency of unchanged positions to infer μt ?

One approach to deal with rate inequality would be to use prior information to divide a DNA sequence into *partitions*, sets of positions, each with the same rate of molecular evolution. For example, we could apply the JC model separately to subsets of the sites that we believe have the same elevated or depressed rate of evolution. The problem with this is that it requires that, before looking at the data, we predict which subsets of sites will have the same rate of evolution. This is difficult to do in most cases.

The alternative is to assume that the rate of substitution, μ , is not the same for all sites, but is drawn from a distribution of rates. By defining a form for the distribution of rates among sites (most commonly a gamma distribution), and by assigning different sites to different rate categories, it is possible to allow for site-to-site rate heterogeneity without having to decide, in advance, which sites are rapidly or slowly evolving. The details are covered by some of the recommended further readings.

In addition to rate heterogeneity, models of molecular evolution can allow for the possibility of nonindependence between nucleotide positions. This might arise, for example, due to base pairing during folding of RNA molecules or the translation of the three bases of each codon into an amino acid. Also, as elaborated in Chapter 11, we can use molecular clock models, which use the same basic substitution models but place constraints on the lengths of branches to force all living species to be the same evolutionary distance from the root.

As the foregoing illustrates, it is possible to develop more and more sophisticated models of evolution to accommodate our knowledge of how DNA sequences actually evolve. The details of these more sophisticated models are less important to grasp than the general principles: we can build realistic models of how DNA sequences evolve and use them to calculate the probability that particular substitutions occur. We can also use the data to guide the selection of an appropriate model of evolution, as discussed later in this chapter.

Up until now we have only considered models of DNA evolution, but phylogeneticists employ many other kinds of data. For many, but not all, of the classes of data described in Appendix 1, continuous-time Markov models have been developed. These include protein sequences, morphology, indels,

restriction fragment length polymorphisms, and amplified fragment length polymorphisms. As a result, model-based methods of phylogenetic inference are now available for almost all widely used data types.

DISTANCE METHODS

The core principle underlying distance methods is that if we knew the true evolutionary distances between each pair of taxa (defined as the average number of substitutions per site in a DNA sequence), then these distances would correspond to only one tree. Because the evolutionary distance between any two taxa is the sum of the lengths (evolutionary distances) of all of the branches on the path between those two taxa, knowing the true evolutionary distances amounts to knowing the tree. For example, if the tree shown in Figure 8.11 were correct, the true evolutionary distances between each pair of taxa would be those given in Table 8.2. If you were given just the distances in Table 8.2, you could work backward to draw an unrooted version of the phylogram in Figure 8.11. The

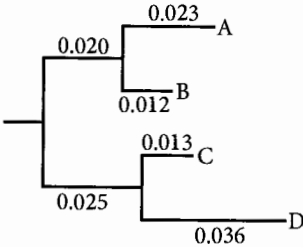


FIGURE 8.11 Phylogram showing evolutionary distances (in substitutions/site).

TABLE 8.2 Evolutionary distances between the taxa in Figure 8.11				
	A	B	C	D
A	0			
B	0.035	0		
C	0.081	0.070	0	
D	0.104	0.093	0.049	0

aim of distance methods is to determine the evolutionary distances between taxa and then use those to infer the true phylogeny.

The starting point for distance-based phylogenetic methods is usually the calculation of the proportion of traits that differ between those taxa, their *pairwise distance*. For this example, Table 8.3 lists the first 10 characters from the carnivoran morphology data set (from Tables 7.2 and 7.4). The pairwise distance is the proportion of characters for which a pair of taxa have a different character state. For example, the outgroup and the cat differ at five characters (2, 4, 7, 8, and 9). We divide this by the total number of characters to get their pairwise distance, 5/10, or 0.5. The pairwise distance matrix for these taxa based on the data in Table 8.3 is shown in Table 8.4.

Table 8.5 provides the pairwise distances using the carnivoran molecular data. This is calculated the same as for morphological data. However, the convention is to only count sites where both taxa have the character scored. Thus, positions that are missing in one or both taxa (including indels) are excluded in the calculation of pairwise distance.

The simple-minded approach to phylogenetic inference would be to take these observed distances and suppose that they are reasonable estimates of the

TABLE 8.3 Ten characters from the carnivoran morphology data set

	Characters									
	1	2	3	4	5	6	7	8	9	10
Outgroup	0	0	0	0	0	0	0	0	0	0
Cat	0	1	0	1	0	0	1	1	1	0
Hyena	0	1	0	1	0	0	1	0	1	0
Civet	0	1	0	0	0	0	0	0	1	0
Dog	1	0	0	0	1	0	0	0	0	0
Raccoon	1	0	0	0	1	0	0	0	0	0
Bear	1	0	0	0	1	1	0	0	0	1
Otter	1	0	0	0	1	0	0	0	0	1
Seal	1	0	1	0	1	1	0	0	0	1
Walrus	1	0	1	0	1	1	0	0	0	1
Sea lion	1	0	1	0	1	1	0	0	0	1

TABLE 8.4 Pairwise distances for the morphological data

	Outgroup	Cat	Hyena	Civet	Dog	Raccoon	Bear	Otter	Seal	Walrus
Cat	0.5									
Hyena	0.4	0.1								
Civet	0.2	0.3	0.2							
Dog	0.2	0.7	0.6	0.4						
Raccoon	0.2	0.7	0.6	0.4	0					
Bear	0.4	0.9	0.8	0.6	0.2					
Otter	0.3	0.8	0.7	0.5	0.1	0.1				
Seal	0.5	1	1	0.7	0.3	0.3	0.1	0.2		
Walrus	0.5	1	1	0.7	0.3	0.3	0.1	0.2	0	
Sea lion	0.5	1	1	0.7	0.3	0.3	0.1	0.2	0	0

TABLE 8.5 Pairwise distances for the carnivoran molecular data

	Mole	Cat	Hyena	Civet	Dog	Raccoon	Bear	Otter	Seal	Walrus
Cat	0.244									
Hyena	0.269	0.092								
Civet	0.246	0.081	0.092							
Dog	0.277	0.190	0.205	0.198						
Raccoon	0.288	0.206	0.212	0.204	0.190					
Bear	0.270	0.175	0.185	0.171	0.179	0.135				
Otter	0.296	0.205	0.209	0.199	0.207	0.121	0.148			
Seal	0.289	0.190	0.196	0.189	0.182	0.151	0.126	0.154		
Walrus	0.283	0.187	0.199	0.183	0.189	0.146	0.128	0.147	0.056	
Sea lion	0.287	0.193	0.198	0.187	0.188	0.147	0.127	0.149	0.058	0.028

true evolutionary distances. Then we could search for a tree that comes closest to predicting this set of evolutionary distances. While it is not uncommon to use this approach, it is inadvisable. As discussed in the context of the card-changing fairy, pairwise distances accumulate more slowly than evolutionary distances because multiple changes occurring at the same site do not always increase the pairwise distance. As a result, the first step in distance analysis is generally to estimate evolutionary distances from observed pairwise distances by correcting the distances based on expectations calculated under a particular model of character evolution.

A pairwise distance can be converted to an evolutionary distance by using the expected relationship between pairwise distance and evolutionary distance. Figure 8.12 shows this for the Jukes-Cantor model of DNA sequence evolution. This graph is the same as Figure 8.4 except that the axes have been relabeled to fit the current context (distance between two taxa rather than frequency of differences between an ancestor and a descendant). Given such a graph, a pairwise distance of 0.24 would be converted into an evolutionary distance of 0.3. Table 8.6 gives the estimated evolutionary distances for the carnivoran molecular data under the Jukes-Cantor model.

Having estimated evolutionary distances, distance methods aim to find the tree that is most consistent with these distances. The first approach is to conduct a series of calculations on the distance matrix that lead directly to a tree. The most widely used method is the *neighbor-joining* (NJ) algorithm (which is described in detail in Swofford et al. 1996 and many online resources). NJ has

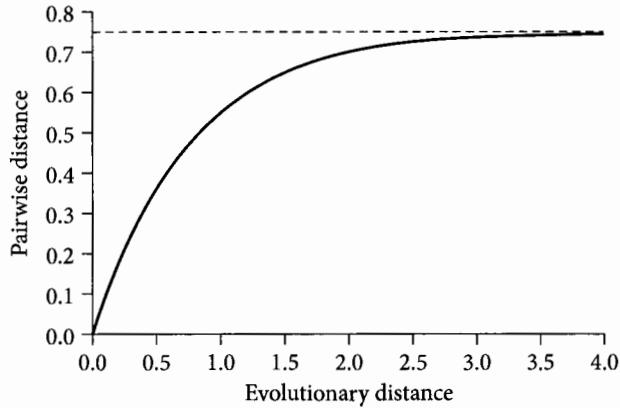


FIGURE 8.12 Relationship of evolutionary and pairwise distances under the JC model. The dashed line indicates the maximum expected pairwise distance, 0.75.

TABLE 8.6 Evolutionary distances estimated under the Jukes-Cantor model for the carnivoran molecular data

	Mole	Cat	Hyena	Civet	Dog	Raccoon	Bear	Otter	Seal	Walrus
Cat	0.300									
Hyena	0.343	0.098								
Civet	0.305	0.086	0.099							
Dog	0.362	0.225	0.244	0.236						
Raccoon	0.371	0.243	0.251	0.239	0.223					
Bear	0.345	0.201	0.215	0.197	0.208	0.150				
Otter	0.389	0.241	0.247	0.233	0.247	0.134	0.166			
Seal	0.376	0.223	0.231	0.222	0.213	0.171	0.140	0.175		
Walrus	0.366	0.218	0.234	0.214	0.222	0.164	0.142	0.166	0.058	
Sea lion	0.373	0.226	0.234	0.219	0.220	0.166	0.140	0.168	0.061	0.029

the virtue that, if the corrected pairwise distances correspond exactly to a tree that would predict those distances, then this tree will be identified by the algorithm. It is also an extremely quick method: even for a very large data set, a tree can be obtained in a fraction of a second. The downside of NJ is that it yields a tree but does not attach a measure of quality to that tree. This means that NJ does not allow us to determine whether a particular tree is significantly better or worse than another. For these reasons, NJ is widely used in situations where we need a quick, but approximate estimate of the true tree. Figure 8.13 shows the neighbor-joining tree for the carnivoran molecular data obtained using the JC distances.

Most other methods for estimating trees from a distance matrix use optimality criteria, that is, measures of tree quality. In parsimony, the optimality criterion is tree length and we search for the tree with the lowest length. For distance methods, a widely used optimality approach is to search for the tree (with branch lengths) that minimizes the difference between the distances in the matrix and the distances that are predicted by the tree. The first step is to calculate the expected evolutionary distances between each pair of taxa on a tree by summing up all of the intervening branch lengths. The set of expected distances is compared to the corrected pairwise distances. One way to quantify

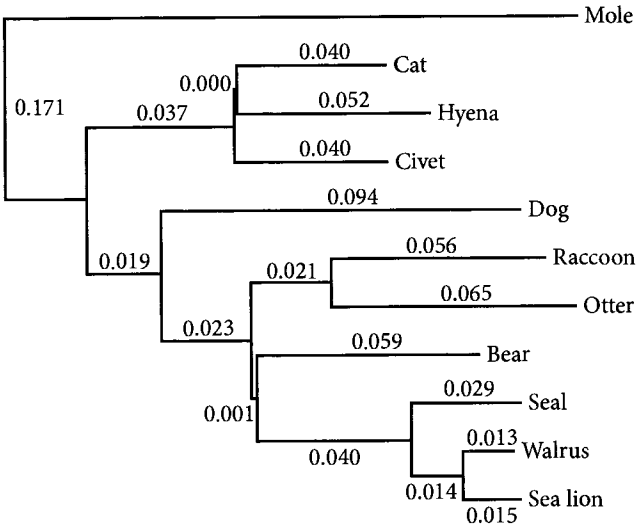


FIGURE 8.13 Neighbor-joining tree based on the JC distances obtained from the carnivoran molecular data. Branch lengths are given in average number of substitutions per site (evolutionary distance).

the fit between a tree and the observed distances is to sum the squared differences between each observed and expected distance. We can then search for a tree topology and a set of branch lengths that minimizes this metric. This tree could be said to be optimal because it comes closest to predicting the observed distances.

There are several variants of this basic approach (see Further Reading). The one we will discuss here is *minimum evolution*. This starts by choosing the optimal branch lengths for a given tree topology using the least squares method: adjusting branch lengths to minimize the sum of the squared deviations between the observed and expected distances. However, rather than picking the tree topology that minimizes this same metric, minimum evolution picks the tree topology on which the total branch length (the sum of the lengths of all the branches) is minimized.

A heuristic search can be conducted to find the optimal tree for a particular distance matrix. These are similar to the searches described for parsimony (Chapter 7) except that the value that is calculated for each tree that is visited is not the parsimony score but the measure of fit between the tree and the distance data (either the squared deviation or the sum of the lengths of all the branches). In comparison to parsimony, which only considers topology when determining the score of a tree, a distance search needs to explore branch length. This explains why distance optimality methods tend to be significantly slower than parsimony for the same number of taxa.

Figure 8.14 shows the distance tree estimated from the carnivoran molecular data using the minimum evolution method. It is worth noting that this tree does not exactly predict the observed distances. For instance, the branches between mole and cat sum to 0.251 ($0.172 + 0.037 + 0.001 + 0.041$) although their estimated evolutionary distance was 0.300 (Table 8.6). This difference could be due to chance events during evolution, errors in the estimation of evolutionary distances from observed pairwise distance, or both. This tree is similar but not identical to the NJ tree, which is expected given that they both used the same evolutionary distance matrix.

When converting character state data to distances, some information is discarded. There is only one distance matrix for a given character state matrix, but many character state matrices can yield the same distance matrix. Because of this loss of information, distance methods tend to have less statistical power than character state matrix approaches, such as parsimony or maximum likelihood.

Minimum evolution and similar optimality methods remain useful in cases in which the original data are already in the form of a pairwise distance (e.g.,

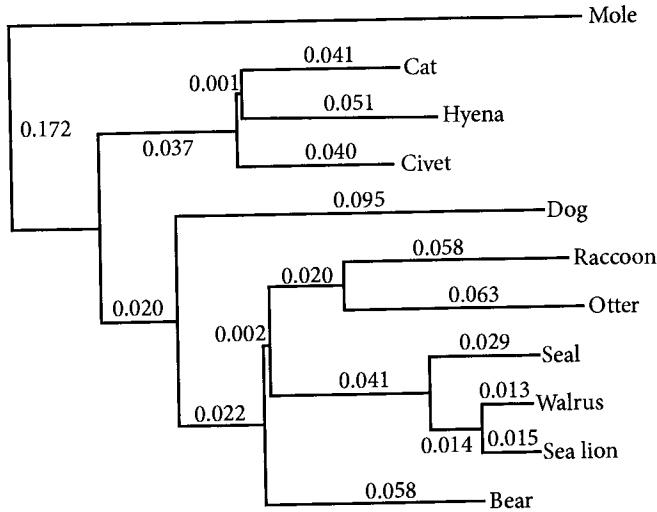


FIGURE 8.14 Minimum evolution tree based on the JC distances obtained from the carnivoran molecular data. Branch lengths are given in average number of character state changes per character.

DNA–DNA hybridization data). Also, it has been found that some specialized distance corrections provide computationally efficient ways to deal with cases where base composition varies among taxa. Nonetheless, you will encounter distance optimality methods much less frequently than neighbor-joining, maximum likelihood, or Bayesian methods.

MAXIMUM LIKELIHOOD

The maximum likelihood (ML) criterion is not specific to phylogenetics, but is a general approach used throughout statistics. Indeed, the early development of parsimony was guided by a desire that it should approximate maximum likelihood. However, the computational complexity of ML delayed its implementation as a method of phylogenetic analysis until the 1990s.

The application of ML to phylogenetics involves searching for the tree that has the highest probability of giving rise to the observed data. Before delving into the application of likelihood to biological data, let us begin by exploring the underlying principles using a coin example.

MAXIMUM LIKELIHOOD

Suppose you have a bag of coins and you know that half of the coins are fair (50% chance of a head) and half of the coins are biased (75% chance of a head). You draw one coin from the bag and wish to consider two alternative hypotheses: the coin is fair versus the coin is biased. The coin is tossed 10 times and each time it falls heads-up. You may now apply likelihood to ask whether the observed data (10 heads) support one of the hypotheses and, furthermore, whether the data are decisive enough to make it reasonable to reject the alternative hypothesis.

The first stage is to specify a model of how coin tossing works. Let us assume that coins all have a head on one side and a tail on the other, that each toss is independent of previous tosses, and that you are 100% accurate in distinguishing heads and tails. Under this model we can calculate the *likelihood*, which is defined as the probability that the data would have arisen under the hypothesis. With 10 heads, the likelihood for a fair coin is 0.5^{10} or ~ 0.00098 (Figure 8.15). This result does *not* mean that there is a 0.1% chance that the coin is fair. It just means that there is a 0.1% chance of this specific outcome for a fair coin.

How probable are the observed data under the hypothesis that the coin is biased? The likelihood under the bias hypothesis is 0.75^{10} or ~ 0.0563 (Figure 8.15). This is still a low number, which tells us that even under the bias hypothesis this particular outcome is improbable. What counts, however, is not the absolute value of the likelihood but a comparison of the likelihoods of the two competing hypotheses. The data are $0.0563/0.00098$, or approximately 56 times as probable under the biased coin hypothesis as under the fair coin hypothesis. This *likelihood ratio*, which is usually presented as a natural logarithm, is a measure of the evidential support for one hypothesis over the other. In this case, the log-likelihood ratio is $\ln(56) = 4.02$. Because 4.02 is well above 2.0, a











Toss	1	2	3	4	5	6	7	8	9	10	Likelihood
Result											
Prob. if fair	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.001
Prob. if biased	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.056

FIGURE 8.15 Likelihood of ten sequential heads for a fair or a biased coin. The likelihood is the product of the probabilities of each individual toss, e.g. 0.5^{10} , under the fair coin hypothesis.

commonly used threshold of significance (in some circumstances, a likelihood ratio of 2.0 approximates the traditional $P < 0.05$ confidence threshold), we would say that the data strongly support the conclusion that the coin is biased.

Now let us consider how likelihood is applied to phylogenetic inference. In this case, the observed data are the characters for each taxon (the character state matrix) and the hypotheses are all the possible trees. Our aim is to determine the probability of the data arising under each tree on the principle that the tree that has the highest likelihood is the best estimate of the true tree. Similar to distance methods, when we talk of a “tree” in a likelihood context, we are thinking of both topology and branch lengths because both factor into our assessment of tree quality.

The first thing we need is a mathematical model of character evolution analogous to the model that specified that the probability of a fair coin coming up heads is 0.5. To illustrate the principles, let us consider the JC model and the simplest possible tree, comprising just two taxa, from which we have obtained a six base-pair DNA sequence (Figure 8.16). The only aspect of this tree that is unknown is the length of the branch separating these two taxa. For four of the positions, both taxa have the same base, indicating that either no change happened or there was a change to a new state and back again. The fact that four out of six bases are identical provides evidence that the branch is not infinitely long because, if it were, we would expect an average of 1.5 matching bases rather than 4.0 (because, under the JC model, $\frac{3}{4}$ of the six bases should differ after an infinite amount of time; Figure 8.12). For two of the positions the two taxa have different bases, showing that the intervening branch is not zero length. The maximum likelihood criterion proposes that the best estimate of the branch length is that which yields the highest likelihood.

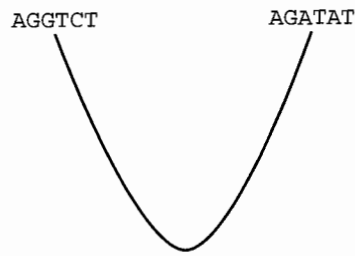


FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips. The only element of uncertainty in this tree is the length of the single branch.

Recall that the branch length is equal to the evolutionary distance between these two taxa (the number of substitutions per site). This will be determined by the branch’s duration and its rate of evolution. These two quantities are hard to disentangle (as will be discussed in the molecular dating section of Chapter 11), but all we need to worry about here is their product: μt . The relationship between branch length and probability of change is given in Figure 8.5.

In both taxa, the first base in the sequence is an A. Considering only one taxon for a moment, the probability that this first base is an A is $\frac{1}{4}$ (since, under Jukes-Cantor, we assume that the bases are at equal frequency). Given that it was an A in the first taxon, the probability that it is an A after μt units is $\frac{1}{4} + \frac{3}{4}e^{-4/3\mu t}$ (see Figure 8.5). So the probability of seeing the data at site 1 is $\frac{1}{4}(\frac{1}{4} + \frac{3}{4}e^{-4/3\mu t})$. We can substitute any value of μt (the branch length) into this equation to obtain the probability that this site would have evolved given this branch length. This probability is called the *site likelihood*. The second site has the same pattern and thus will have the same site likelihood for any branch length. The third site has a G in one taxon (probability $\frac{1}{4}$) and an A in the other taxon. The probability of this substitution is $\frac{1}{4} - \frac{1}{4}e^{-4/3\mu t}$, giving character three a site likelihood of $\frac{1}{4}(\frac{1}{4} - \frac{1}{4}e^{-4/3\mu t})$.

The likelihood of a tree knowing the entire character matrix is given by the product of the individual site likelihoods. The likelihood is based on a product, rather than a sum, because all of the characters need to attain their observed states in order to obtain the full data matrix. When we multiply the individual site likelihoods, we obtain an overall likelihood of $[\frac{1}{4}(\frac{1}{4} + \frac{3}{4}e^{-4/3\mu t})]^4 [\frac{1}{4}(\frac{1}{4} - \frac{1}{4}e^{-4/3\mu t})]^2$. The first element, raised to the power of four, refers to the four sites that are unchanged, and the second element, raised to the power of two, refers to the two sites that differ between the taxa. Using this formula, we can then consider the likelihood that different branch lengths (values of μt) gave rise to the data in Figure 8.16. Looking at Figure 8.17, we observe that a branch length of $\mu t = 0.44$ has the highest likelihood (0.595×10^{-6}). This shows us that, under the JC model, 0.44 is the best estimate of the length of the branch in Figure 8.16.

Instead of keeping track of the likelihood, it is conventional to record the natural logarithm of the likelihood, the *log-likelihood*. Using logarithms is helpful to avoid computer problems associated with handling very small numbers. A likelihood of 0.595×10^{-6} corresponds to a log-likelihood of -14.33 ($e^{-14.33} = 0.595 \times 10^{-6}$). The objective of maximum likelihood analysis is to find the tree with the highest likelihood. This corresponds to the least negative log-likelihood, which is -14.33 for the data in Figure 8.16.

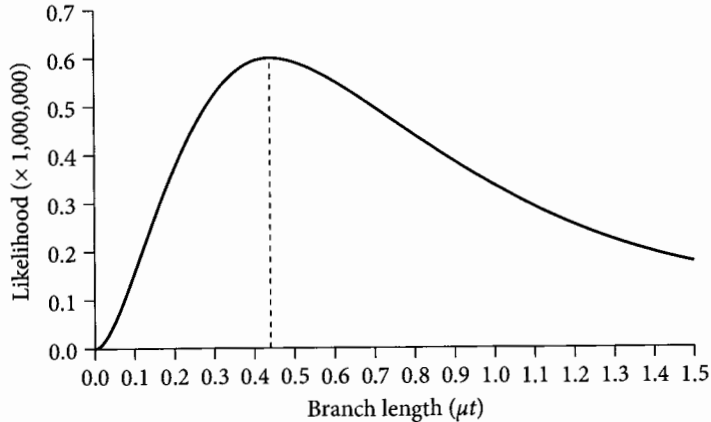


FIGURE 8.17 Likelihood values for different branch lengths given the data shown in Figure 8.16.

To step back, we have just used a model of molecular evolution to find the tree that has the highest probability of generating the observed data. This does not mean that this tree is true. It just means that, if the model of evolution is correct, this is the best point estimate of the tree based on the current data. Furthermore, we can say that as we move away from the optimal tree, the likelihood becomes lower, meaning that the observed data are successively less probable.

You may have wondered why we worked with such a simple example: a tree composed of just two taxa and the JC model of molecular evolution. The reason is that likelihood calculations rapidly become much more complicated as we add parameters to our model. Suppose we selected a slightly more complex model of evolution, for example, F81, which allows bases to have unequal frequencies. This adds three parameters: the frequencies of three of the bases (the fourth base is “free” because we can calculate it by subtracting the others from 1.0). Instead of a likelihood function that can be depicted in a two-dimensional graph (Figure 8.16), we simultaneously need to consider values of four parameters: the three base frequencies and the branch length. This poses a challenge because we need to explore variation in all four dimensions simultaneously to find the set of parameters that maximize the likelihood. Computer scientists have worked out algorithms for doing this efficiently. Basically, a computer program iteratively visits each of the parameters and assigns it a value. Then it cycles among them, sliding the values up and down in an attempt to maximize

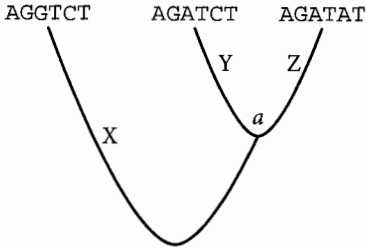


FIGURE 8.18 Three-taxon tree with a six base-pair sequence at each tip. The only items of uncertainty are the three branch lengths (X, Y, and Z) and the sequence at internal node *a*. We use the maximum likelihood criterion to estimate the value of the branch lengths, while summing over all possible sequences at node *a*.

the likelihood. Eventually, the program finds a set of values that cannot be significantly improved upon, and concludes the search.

The problem is even more challenging with larger numbers of taxa, because adding taxa not only adds branches whose lengths need to be considered but also adds ancestors whose character states are unknown and have to be inferred. For example, let us expand the previous example to add a third taxon (Figure 8.18). We now have three branches (X, Y, and Z) instead of just one, and we have created a node, *a*, whose sequence is unknown. The calculation of the likelihood looks at a branch and asks, What is the probability of the data that are observed at the two ends? But what if you have only observed data at one end of a branch?

Although we cannot observe the identity of the base at node *a* for any of the six characters, for each character we know that it was one of the four bases, A, C, G, or T. This means that we can determine the site likelihoods by summing over the four possible states at node *a*. We take the sum, rather than the product, because the observed data could have evolved because node *a* had state A or C or G or T. However, as before, we multiply the site likelihoods to obtain the overall likelihood. Under the Jukes-Cantor model, the maximum likelihood estimates of the lengths of the three branches in Figure 8.18 are $X = 0.188$, $Y = 0$, $Z = 0.188$. The log-likelihood of this tree is -15.92 .

The computational challenges get even harder as we add additional taxa. With four taxa we have two internal nodes and, thus, 4^2 (16) possible sets of ancestral states that we need to sum over to determine the site’s likelihood. With five taxa there are 4^3 (64) histories. In general, there are $4^{(n-2)}$ histories, where n is the number of taxa in the tree. The necessity of summing over all

these histories is one of the principal reasons that maximum likelihood analysis is so computationally intensive. Fortunately, advances in computation and some creative shortcuts have made these calculations practical even for very large data sets.

To recap: computer programs that perform phylogenetic analysis using maximum likelihood follow four steps. First, a particular tree and parameters (including branch lengths) are set and the likelihood of each site (character) is determined by summing over all possible histories for that site. Second, the likelihoods of each site in the matrix are multiplied (their logarithms are added) to obtain the overall likelihood of *that* tree with *those* parameters and branch lengths. Third, the program optimizes the branch lengths and other parameters by changing them iteratively and repeating the first two steps until the likelihood is maximized. This gives the maximum likelihood estimates of these parameters for the first tree. Fourth, a search through tree space is conducted to find the maximum likelihood tree, the tree on which the probability of the data is maximized. For every tree considered in the last stage in this process it is necessary to propose a set of branch lengths, calculate the likelihood, and then iteratively optimize the branch lengths. This should yield the tree that has the highest likelihood.

It is probably apparent that the computational challenge of maximum likelihood is great even for a simple model of molecular evolution. This is why there was such a lag between the 1970s, when the maximum likelihood method was first applied to phylogenetic inference (Felsenstein 1973), and the mid-1990s, when it became feasible to apply the method to real data sets. In the intervening years, computers became much faster and computer scientists and theoreticians found some effective shortcuts for doing likelihood calculations. As a result, it is now possible to do a full likelihood analysis of a data set like the carnivorans in less than an hour on a personal computer. The maximum likelihood estimate of the optimal tree for the carnivoran molecular data (using the HKY model of evolution) has a log-likelihood of approximately -5184. As shown in Figure 8.19, this tree is similar to the neighbor-joining tree but differs in the position of the bear and hyena lineages.

Currently available models of morphological evolution are rather simplistic. For example, they typically assume that all characters have the same rate of evolution and that the only character states a character can adopt are those that were observed in at least one of the taxa. Nonetheless, it is still possible to apply maximum likelihood to morphological data. Figure 8.20 shows the results of a maximum likelihood analysis of the carnivoran morphological data. This tree has an identical topology to one of the two maximum parsimony trees.

MAXIMUM LIKELIHOOD

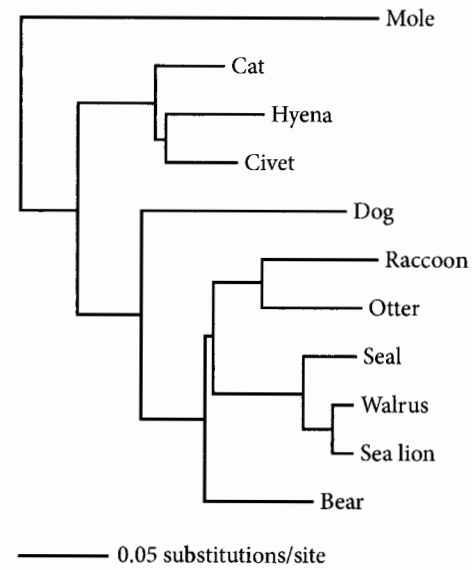


FIGURE 8.19 Maximum likelihood tree based on the carnivoran molecular data. A scale bar is provided to indicate branch length.

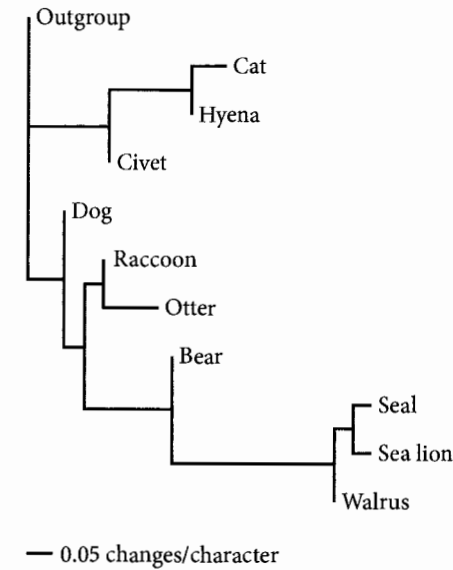


FIGURE 8.20 Maximum likelihood tree based on the carnivoran morphological data. A scale bar is provided to indicate branch length.

CHOOSING AN EVOLUTIONARY MODEL

As we have seen, the maximum likelihood criterion provides a way to estimate the value of any parameter in a model. Of particular note, base frequencies and the relative rates of different kinds of substitutions are estimated from the data at the same time as a tree is being inferred. This differentiates model-based methods from generalized parsimony (Chapter 7), which lacks any objective way to select the right costs to apply to different kinds of changes. Generalized parsimony asks the practitioner to assert whether there is a transition:transversion bias and, if so, how much of a cost should be used to reflect this fact. In contrast, under maximum likelihood, the data themselves are used to select the parameter values that maximize the probability of obtaining the observed data.

While maximum likelihood picks the parameters of the model (e.g., the base frequencies in F81, the transition:transversion bias in HKY), how do we pick the class of model to use? How do we decide whether to use JC, F81, HKY, or GTR? Likewise, how do we decide whether or not to allow for rate variation across sites and whether a molecular clock applies? It turns out that we can actually use the data to guide the choice of models. To get a feel for how this works, it is important to understand how the likelihood will change as you add more parameters to a model.

As discussed earlier, models with fewer parameters can be understood to be special cases of parameter-rich models (see Figure 8.10). The simpler model is said to be *nested* within the more complex one. When this is the case, the more complex (parameter-rich) model is *guaranteed* to have a likelihood that is equal to or greater than the simpler model. This makes sense. Each parameter added to a model makes the model more able to adapt to features of the data. In the same way that a metal chain with many smaller links can wrap itself more precisely around a pole than a chain with a few large links, a model with more free parameters can describe the data more precisely and will therefore yield a higher likelihood than a simpler model.

You might think that because more complex models always yield higher likelihoods, we would always use the most complex model we can define. However, there are costs to using an excessively complex model. A more complex model has more parameters and inevitably slows down the analysis. Also, in much the same way that a chain of a given length with more links is weaker than one with fewer links, more complex models come at the cost of reduced

statistical power. It may be impossible to say if tree 1 is significantly better than tree 2 when using an overly parameter-rich model, while a more appropriate model might allow us to conclude with confidence that tree 2 is false. When it comes to parameters, we *can* have too much of a good thing.

If there are costs to overly simple models (inability to account for the processes underlying the data) and to overly complex ones (losing the power to distinguish models), how do we choose which model to use? The principle is simple enough: pick a more complex model only when the gain in likelihood is more than would be expected if the simpler model were true. For example, if evolution followed the JC model, we would get a higher log-likelihood when we assumed the F81 model, but it should not be *much* higher than if we had assumed JC. If it were, we would have grounds to suspect that the underlying base frequencies really were different from 25%.

Statisticians have developed ways to predict how much higher the log-likelihood would be under a more complex model if the simpler were true. Different approaches use different formulae for calculating the expected likelihood gain due to adding extra parameters. One common approach (*hierarchical likelihood ratio tests*) can be used to compare nested models, where the more complex model has p extra parameters. Readers with a background in statistics may be interested to know that twice the log-likelihood difference between the two models is expected to fit a chi-square distribution with p degrees of freedom. For example, recalling that the F81 model has three more free parameters than the JC model (Figure 8.10), there is only a 5% probability that the F81 model will be more than 3.9 log-likelihood units higher than the JC model if the JC model is in fact true. This is because, under the χ^2 distribution, a value of ~ 7.8 ($= 3.9 \times 2$) corresponds to a P -value of 0.05 for 3 degrees of freedom. Thus, if for a particular data set F81 yields a log-likelihood that is 4 or more units higher than JC, then we can be confident that the assumption of equal base frequencies is violated.

BAYESIAN INFERENCE

The newest approach to estimating phylogenies is Bayesian inference. Whereas likelihood judges a tree based on how probable it is that evolution would have produced the observed data, Bayesian inference judges trees based on their posterior probability, the probability that the tree is true, given the data, our

models of evolution, and our prior beliefs. To give you a feel for the Bayesian approach, let us return to the coin example introduced in the last section.

Recall that you were given a bag of coins, half of which are fair (50% heads) and half of which are biased (75% heads). You pull a coin out of the bag and want to know if it is biased or not. Before even tossing the coin, the *prior probability* of it being biased is 0.5 because we know that half the coins in the bag are biased. After tossing the coin, we can use the results to update this probability. Because this extra information comes after we have collected data, it is called the *posterior probability*.

Recall that we were able to deduce previously that having a biased coin was more likely to generate the observed outcome than a fair coin, but we did not actually calculate the probability that the coin was biased. The principles for calculating such posterior probabilities were developed by the Reverend Thomas Bayes in the 18th century. He proved mathematically that the probability of a hypothesis, given some data, is equal to the probability of the data, given the hypothesis (the likelihood), times the prior probability of the hypothesis and divided by the probability of the data (summed over all hypotheses). Or more formally:

Posterior
probability

Likelihood

Prior probability
of the hypothesis

↓

↓

↓

Bayes' theorem: $\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D)}$

← Prior probability
of the data

In this equation, Pr refers to “probability,” D to “data,” and H to “hypothesis.” The vertical line is read as “given.” For example, $\Pr(H|D)$ should be read as “the probability of the hypothesis, given the data.”

We can apply this equation to obtain the posterior probability of the coin being biased. The probability of the data (10 heads), given the hypothesis that the coin is biased, $\Pr(D|H)$, is the likelihood. Using the same model as previously, this is 0.75^{10} , or 0.0563. The prior probability, $\Pr(H)$, that the coin is biased is 0.5. So the numerator in this example is 0.0281.

What is the prior probability of the data, the denominator in Bayes’ equation? You might imagine that this probability is 1.0, seeing as the coin must be either fair or biased. But remember that we are trying to determine the probability of getting 10 heads in 10 tosses, as opposed to 9 heads and 1 tail, 8 heads and 2 tails, and so on. This probability will certainly be less than 1.0.

To calculate the prior probability of the data, $\Pr(D)$, we need to determine the probability of obtaining those data under each hypothesis and then we need to sum over all possible hypotheses (weighted by the hypotheses’ prior probability). Here, there are only two possible hypotheses, namely, that the coin is biased or that the coin is fair. There is a 0.5 chance it is fair, and if it is fair the probability of getting 10 heads is 0.5^{10} . There is a 0.5 chance that it is biased, and if it is biased the probability of getting 10 heads is 0.75^{10} . Summing these together, the probability of the data is $0.5 \times (0.5^{10} + 0.75^{10}) = 0.0286$. This means there was a 2.9% chance that you would have grabbed a coin at random from the bag and then obtained 10 heads in 10 tosses.

Combining these numbers, the posterior probability that the coin is biased after observing 10 heads in a row is $0.0281/0.0286$, or about 0.98. This means that, given the priors and model, there is a 98% chance that the coin is biased and only a 2% chance that it is fair. By taking into account the observations of 10 consecutive heads, the probability that you drew a biased coin has jumped from 0.5 to 0.98, while the probability that you drew a fair coin has dropped from 0.5 to 0.02.

What is interesting and special about the Bayesian approach is that the starting information matters. Suppose, for example, that you drew the coin from a sack that had only 1% biased coins. In that case the posterior probability that it is biased after getting 10 heads in a row is only 0.37. While the data has moved you from a posterior probability of 0.01 to 0.37, the posterior is still less than 0.5. This means that even after observing 10 heads you would still be wise to bet against it being a biased coin!

In this coin example, the alternative hypotheses each indicated an exact probability of heads, and hence a simple calculation of the likelihood of the data. What if you knew that the sack contained two kinds of coins at equal frequency: “fair” coins whose probability of yielding heads is between 0.4 and 0.6 and biased coins whose probability of heads is somewhere between 0.7 and 0.9? What would you do? By extrapolation from the discussion of maximum likelihood, you might argue for selecting whatever value within these ranges maximizes the likelihood. Thus, if you observed 10 heads you would select a value of 0.6 for the fair coin hypothesis (likelihood = 0.6^{10}) and 0.9 for the biased coin hypothesis (likelihood = 0.9^{10}). However, this is *not* what the Bayesian approach calls for. Instead, a good Bayesian would integrate over all possible values of the parameter, weighted by the prior probability of each being the true value of the parameter. This is one reason why Bayesian methods can

yield different conclusions than maximum likelihood, even in cases where all hypotheses have the same prior probability.

Now let's apply these principles to phylogenetics. The data correspond to a character state matrix and the hypotheses correspond to the alternative possible tree topologies. Thus Bayes' theorem takes the following form:

$$\Pr(\text{Tree}|\text{Data}) = \frac{\Pr(\text{Data}|\text{Tree}) \times \Pr(\text{Tree})}{\Pr(\text{Data})}$$

The prior probability of a particular tree topology, $\Pr(\text{Tree})$, is the probability (before looking at your data) that among all possible trees it is the true tree. For example, if we believed that all tree topologies were equally likely *a priori*, we could apply a *flat prior*, where the prior probability of each tree equals one divided by the number of distinct tree topologies (see Table 7.8).

Calculating the probability of the data given the tree, $\Pr(\text{Data}|\text{Tree})$, entails determining the likelihood of the tree. This is done as described earlier in this chapter, except that instead of selecting the values of the free parameters (e.g., base frequencies, branch lengths) that maximize the likelihood of the data, we integrate over the prior probability for all parameters. These are often difficult calculations, but they are possible in many cases.

The big challenge with Bayesian phylogenetics is calculating the prior probability of the data, $\Pr(\text{Data})$. The problem is that $\Pr(\text{Data})$ involves a summation over all trees, but, as shown in Table 7.8, there are lots of possible tree topologies. Whereas a single tree in isolation can be assigned a parsimony score or likelihood, a Bayesian posterior probability cannot be assigned to a single tree without taking account of all other possible trees. As a result, Bayesian phylogenetics would be impossible without the invention of a clever method called *Markov chain Monte Carlo* (MCMC) analysis.

The MCMC method exploits the fact that while we cannot easily calculate the actual posterior, we can calculate the relative posteriors of different trees. To help you visualize this, imagine that you want to survey the altitude of every point in a landscape and you know that the true altitudes, summed over all points, is 100,000 m (this is analogous to knowing that the sum of posterior probabilities of all trees must sum to 1.0). You have a defective altimeter—it accurately measures altitudes except that each measurement is multiplied by some unknown constant. If you get a reading of 400 m at one point, you have no idea whether the real height is 40 m or 4000 m. This device is not totally use-

less, however. If you measured a second point at 360 m, you would know that it was 10% lower than the first point (the ratio of the two heights is 10:9). Thus, if you surveyed the entire landscape with the faulty altimeter, you could recalibrate all the measurements so that they summed to 100,000 m.

When applying MCMC to phylogenetic inference, the landscape is a highly multidimensional parameter space. In addition to containing all possible tree topologies, this space includes the range of possible branch lengths and all the free parameters of the model of evolution. This is a much more complicated space than the tree space introduced in Chapter 7, but the principle is identical. The process goes on within a computer, but we can imagine walking through this space ourselves. Figure 8.21 provides a visual representation of an MCMC run.

We start at some point in this parameter space and calculate the likelihood, $\Pr(D|H)$, the probability that evolution would have yielded exactly these data. Given our model of evolution and the parameters specified by our location in parameter space, we can calculate relatively easily the likelihood associated with this point in the landscape.

The next step is to randomly propose a new parameter combination and calculate its posterior probability. We also need to calculate something called the

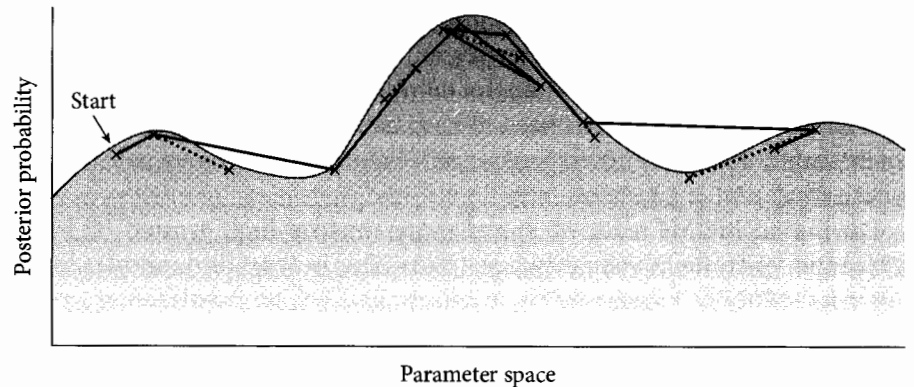


FIGURE 8.21 A visual representation of the Markov chain Monte Carlo method. Each × marks a parameter value that was proposed during the chain. Those proposals that were accepted are marked with a solid line, whereas those proposals that were rejected are marked with a dotted line. All “uphill” proposals are accepted. Some downhill proposals are rejected and some are accepted.

proposal ratio, but we will not describe this ratio (see Further Reading). The rule we will follow is that if the new point in parameter space is uphill, meaning it has a higher posterior probability, we will move to that new point. This higher posterior probability could arise because the new parameter combination yields a higher likelihood and/or because it has a higher prior probability. Either way, analogous to heuristic searching for optimal trees, we always accept a proposal that has a higher posterior probability (Chapter 7).

The difference between an MCMC run and a heuristic search algorithm is that when a proposal takes us downhill (i.e., when the product of the likelihood and prior is lower than the current state), we might nonetheless accept the proposal. We decide randomly (hence the Monte Carlo reference) whether to accept the “downhill” tree/parameter. The probability of accepting such a proposal is based on the ratio of the two posteriors. If there were a 10% difference between the posteriors, meaning the ratio is 10:9, there would be a 9/10 chance of accepting the proposal and a 1/10 chance of rejecting the proposal and thereby staying at the previous parameter value. The process of proposing a new parameter value and deciding whether to accept it is considered one *MCMC generation*.

Starting from whichever parameter combination we chose in the last generation, we then initiate a new generation. A new proposal is made and the same rule is followed to decide whether the proposal should be accepted or rejected. The MCMC run continues for millions of generations, each consisting of a proposal that is either accepted or rejected. During this process, the computer keeps a list of the trees and parameter combinations that were visited during the chain (actually, we do not need to keep data for every generation but only a subsample, e.g., every 100th generation). Figure 8.21 may give you a feel for what a small part of an MCMC run “looks” like.

During an MCMC run, we spend more time on high ground (at better trees) and less time in valleys (worse trees). This is because it takes us longer to walk downhill (because we often reject downhill proposals) than to walk uphill (because we always accept uphill steps). It turns out that the frequency with which we find ourselves in a region of parameter space will eventually be proportional to that region’s posterior probability. It can be mathematically proven that if we wander around the parameter space long enough, the proportion of trees in our list having a specific topology will be proportional to that topology’s posterior probability. Similarly, for all parameters of our model (e.g., branch lengths, transition:transversion bias, rate heterogeneity across sites

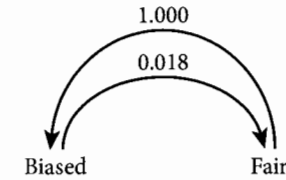


FIGURE 8.22 Representation of a Bayesian MCMC analysis in the coin example. The parameter space has two points, “Fair” and “Biased,” corresponding to the two hypotheses. When the chain is on “Fair,” the proposal of a switch to “Biased” will always be accepted. When the chain is on “Biased,” there is a probability of only 0.018 that the proposal of a switch to “Fair” will be accepted.

in the sequence, etc.) the list of values visited during the MCMC run should approximate their posterior probability distribution.

To illustrate the idea, let us revisit the coin example and see if an MCMC approach will yield the same posterior probability as we obtained by calculation. Imagine two spots on the floor, one representing the case of a fair coin and one representing the case of a biased coin (with a 0.75 probability of yielding heads). We can use MCMC to calculate the posterior probability of the two hypotheses given that they each have a prior probability of 0.5 (Figure 8.22). We will start by stepping on one of the two spots. Then we will “propose” the other spot. The likelihood is higher under the biased coin hypothesis (−0.056) than under the fair coin hypothesis (−0.001). As a result, whenever we are on the spot representing the fair coin hypothesis, the proposal to switch to the other spot (representing a biased coin) will be accepted. In contrast, if we are on the biased coin spot, the probability of accepting the fair coin spot is defined by the ratio of the two likelihoods, meaning that the probability of accepting this proposal is $0.001/0.056 = 0.018$ (Figure 8.22).

Imagine using MCMC to guide a decision as to whether to jump between these two spots. When you are on the biased spot, you will reject the proposal to move to the fair spot ~98% of the time and accept it ~2% of the time. If you accept the proposal, you will spend exactly one generation on the “fair” spot before jumping back to the “biased” spot. It should be obvious that after many generations, you will have spent a total of 98% of the time on the biased spot and 2% on the fair spot. As you will see, these values correspond to the posterior probabilities that we calculated earlier.

Bayesian phylogenetics using MCMC involves several complications, some of which are worth mentioning because they influence the performance of the method and the reliability of the output. First, it is necessary to select a model of trait evolution. Analogous to maximum likelihood analysis, the model may be selected prior to the MCMC analysis using hierarchical likelihood ratio tests and similar methods. Alternatively, it is possible to place prior probabilities on different models and allow the MCMC to jump between models, thereby integrating over uncertainty in model choice.

The second complication with MCMC is that the chain needs to reach *stationarity* before it provides useful information on the posterior probability. In this context, stationarity means that the likelihood is bouncing around but not showing any consistent upward trend. If we start at a very poor set of parameter values, it may take a long time to find our way to high ground, and we would not want the initial period spent at parameter values with low posterior probabilities to distort our conclusions. The general procedure is to identify the period before stationarity has been reached, the so-called burn-in, and delete these entries from the list. This is often done by looking at a plot of likelihood as a function of generation and deleting all generations before the point at which the curve levels out. Figure 8.23 shows an example of such a plot.

A third issue is to ensure that, during the period of stationarity, the run is exploring all of parameter space—that it is *mixing*. If chains are mixing prop-

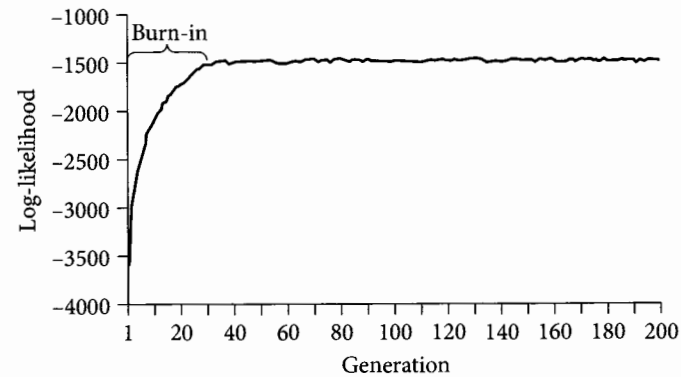


FIGURE 8.23 Change in likelihood during a Bayesian MCMC. Samples collected during the first part of the chain, when there is a steady increase in the likelihood, the burn-in, should be discarded before the posterior sample is summarized.

erly, then different runs will converge on a similar posterior distribution. This can be checked by initiating multiple runs from different random starting points and seeing if each run supports the same phylogenetic conclusions. If we find evidence of poor mixing, the MCMC procedure can be adjusted to improve its performance. Most adjustments include modifying the way that new points in parameter space are proposed. For example, sometimes each run will entail several coupled chains, where only one (the “cold” chain) is following the MCMC rules. In such a strategy, the other (“heated”) chains follow modified rules of decision making and serve as a source of points in parameter space that can be proposed to the cold chain. More information can be found in the recommended Further Reading.

Once we are convinced that we have adequate mixing and thus a reasonable sample of the posterior probability distribution, we can query it to learn about particular parameters, for example, tree topology. By counting how many times a particular tree topology is sampled in the post-burn-in trees, we can obtain an estimate of that tree’s posterior probability. For example, if our distribution contained 8000 trees of which 7200 have the same topology, that tree’s posterior probability would be estimated to be $7200/8000 = 0.90$.

While Bayesian phylogenetics is complex and computationally demanding, the method has become relatively easy and quick to implement thanks to the development of user-friendly programs, such as MrBayes. This program can be run on various kinds of computers and is very flexible. The book *Phylogenetic Trees Made Easy* offers an introduction to the hands-on aspects of using this program for phylogenetic research (Hall 2011).

Analysis of the carnivoran molecular data set using MrBayes yields a posterior distribution containing nine distinct topologies. The tree with the highest posterior probability (0.76) is shown in Figure 8.24. The posterior probability of 0.76 means that, given our prior assumptions about the probability of different trees and models of molecular evolution (which we are glossing over), there is a 76% chance that this tree is correct and a 24% chance that it is wrong in some way.

The morphological data are less supportive of any one topology. The tree that has the highest posterior probability is shown in Figure 8.25. It has a posterior probability of only 7.9%. This low number is probably due to the fact that the morphological data set is small relative to the number of possible trees. With little information, tree space becomes flatter and the MCMC wanders more, spending less time visiting any particular tree.

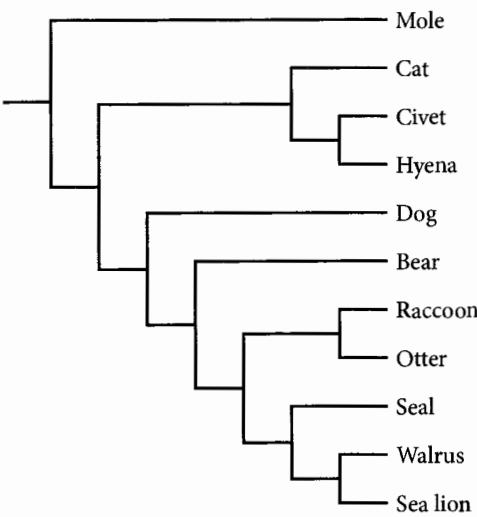


FIGURE 8.24 Tree topology with the highest posterior probability from a Bayesian MCMC analysis of the carnivorous molecular data. This analysis used the GTR+ Γ model of evolution.

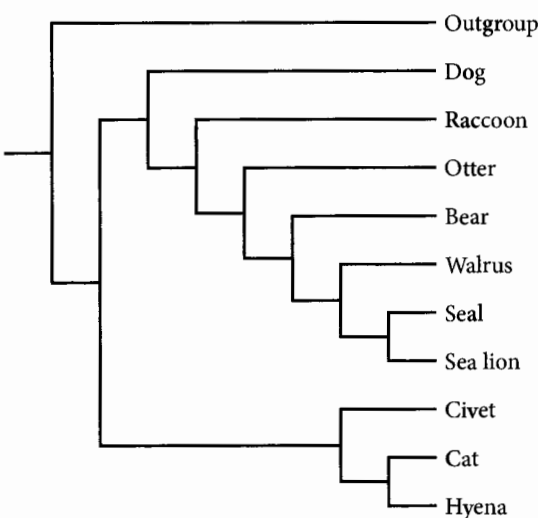


FIGURE 8.25 Tree topology with the highest posterior probability from a Bayesian MCMC analysis of the carnivorous morphological data.

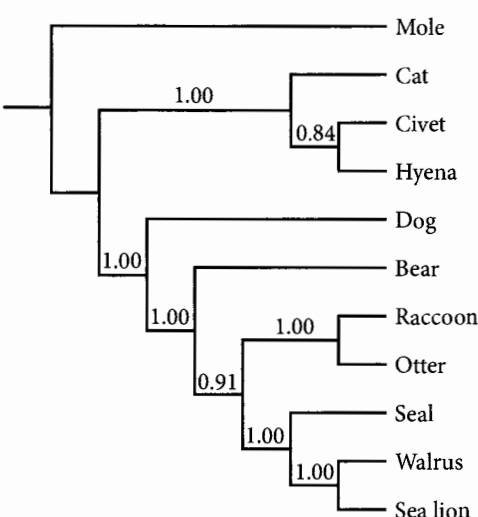


FIGURE 8.26 Bayesian majority-rule consensus tree for the carnivorous molecular data. Numbers on branches are clade posterior probabilities: the fraction of trees in the posterior sample that contain the clade in question. This analysis used the GTR+ Γ model of evolution.

Generally, in phylogenetic analysis, we are not specifically interested in the posterior probability of a particular tree topology. Rather, we care about the posterior probability of individual clades. We can calculate the posterior probability of a particular clade by seeing how often it appears during the MCMC analysis. For example, if a clade is present in 7600 of 8000 trees, then its posterior probability (sometimes called its *clade credibility*) is 0.95. Typically, a posterior distribution is summarized by drawing a tree composed of clades whose posterior probability is greater than 0.5. This is the *Bayesian majority-rule consensus tree*, often loosely called a *Bayesian tree*.

The Bayesian majority-rule consensus tree for the carnivorous molecular data obtained with MrBayes is shown in Figure 8.26. It has the same topology as the single tree with the highest posterior probability. The numbers on each branch are clade credibility estimates. This analysis suggests that all but two clades have a greater than 95% probability of being true (given the model and priors).

The Bayesian consensus tree for the morphological data is shown in Figure 8.27. In contrast to the molecular tree, only three clades have credibility scores greater than or equal to 0.95. Thus, once again we see that the carnivorous

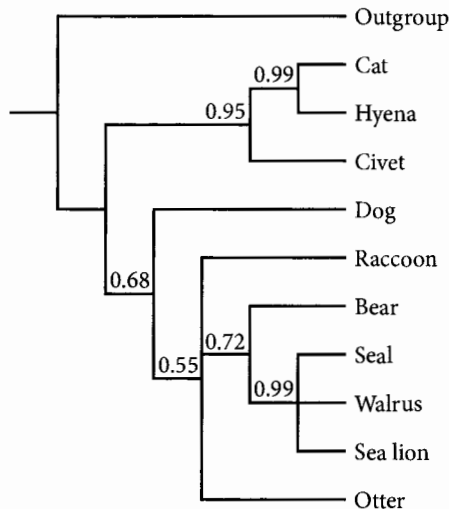


FIGURE 8.27 Bayesian majority-rule consensus tree for the carnivoran morphological data. Numbers on branches are clade posterior probabilities. Branches with posterior probabilities < 0.50 have been collapsed into polytomies.

molecular data set contains more phylogenetic signal than does the morphological data set.

Bayesian phylogenetic analysis has a lot of desirable features, not least of which is that the endpoint is something that scientists usually want, namely, an estimate of the probability that a tree is true (given the data, model, and priors). Nonetheless, some practitioners are uncomfortable with the approach, primarily because of doubts over the very structure of Bayesian statistics. In particular, some phylogeneticists object to having to specify an exact prior probability distribution for all parameters. This, they feel, makes the approach subjective—if you and I have a different set of prior beliefs, the posterior probabilities we obtain will be different. However, one can argue that all of science involves updating probabilities of hypotheses by combining prior knowledge with new data. When we conduct any experiment, we apply some prior knowledge about the system to guide the collection and interpretation of new data. Bayesian inference simply provides a formal way to combine prior knowledge with new information. Given the computational tools and the philosophical appeal of Bayesian inference, it is perhaps not surprising that Bayesian inference is becoming the most widely used method for phylogenetic analysis.

FURTHER READING

General resources: Swofford et al. 1996; Felsenstein 2004
Models of evolution: Swofford et al. 1996; Lewis 1998, 2001
Distance methods: Fitch and Margoliash 1967; Felsenstein 1984; Saitou and Nei 1987; Rzhetsky and Nei 1992
Maximum likelihood: Felsenstein 1981a; Huelsenbeck and Crandall 1997; Lewis 1998
Model selection: Goldman 1993; Posada and Crandall 2001
Bayesian phylogenetics: Larget and Simon 1999; Mau et al. 1999; Huelsenbeck et al. 2002; Holder and Lewis 2003; Huelsenbeck et al. 2004; Lewis et al. 2005

CHAPTER 8 QUIZ

1. In the substitution probability matrix for JC (Figure 8.5), the top right entry is $\frac{1}{4} - \frac{1}{4}e^{-4/3\mu t}$. What does this mean? (μ is the substitution rate; t is time)
 - a. The rate of going from A to T is $\frac{1}{4} - \frac{1}{4}e^{-4/3\mu}$ changes per unit time, t .
 - b. The proportion of changes that occur during a time window that are from A to T is $\frac{1}{4} - \frac{1}{4}e^{-4/3\mu t}$.
 - c. The probability of starting at A and ending at T at each site in the sequence is $\frac{1}{4} - \frac{1}{4}e^{-4/3\mu t}$.
 - d. Answers a and b are correct.
 - e. Answers a, b, and c are correct.

Questions 2–3. Assume that a gene evolves according to the JC model with a substitution rate of 5.2×10^{-10} substitutions per site per year. (*Hint:* Refer to Figure 8.5 and use a calculator or spreadsheet.)

2. A particular site is an A now. What is the probability that this base will be a G after 25 million years?
 - a. 0.9871 b. 0.0617 c. 0.0129 d. 0.0043 e. 0.0011
3. A particular site is an A in species X and a G in species Y. What is the site likelihood under the assumption that their last common ancestor lived 12.5 million years ago?
 - a. 0.9871 b. 0.0617 c. 0.0129 d. 0.0043 e. 0.0011