Chapter 6

Probabilistic Models of DNA Mutation

The methods developed so far have not required any detailed mathematical description of the mutation processes DNA undergoes from generation to generation. In fact, the only real consideration we've given to the mutation process is to point out that if we believe mutations are sufficiently rare events, then both parsimony and use of the uncorrected p-distance are justifiable approaches to inferring a tree. Under such circumstances mutations are unlikely to occur multiple times at the same site, so possible hidden mutations would be negligible.

However, if mutations are not so rare, either because of a high mutation rate or because of a long elapsed time, then along any edge of a tree a particular site might experience more than one change. In the idealized situation of observing both the ancestral and descendant sequence, at most one mutation is observable at a site, though several might have occurred. Under these circumstances both the principal of parsimony and the Hamming dissimilarity map would give inappropriate views as to how much mutation has occurred — they both underestimate the true amount of change.

To do better, we need explicit mathematical models describing how base substitutions occur. Since mutations appear to be random events, we formulate these probabilistically. In subsequent chapters these models will be used first to develop improved dissimilarity maps for distance methods, and then as a basis for the primary statistical approaches to phylogenetic inference, Maximum Likelihood and Bayesian analysis.

6.1 A first example

To begin, we present a simple example in order to illustrate the basic modeling approach. We'll keep this discussion as informal as possible, and then be more careful with our terminology later on. For those who have seen Markov models

before, either in a probability or linear algebra course, the framework will be familiar.

Our model will first describe *one site* in a DNA sequence, and how it changes from an ancestral sequence to a descendant sequence along a *single edge* of a tree. To further simplify, we focus not on the precise base at that site, but only on whether a purine `R` or a pyrimidine `Y` appears.

To be concrete about the situation we wish to describe, suppose we somehow had access to data of an ancestral sequence S0 and a descendant sequence S1:

S0 : `RRYRYRYYRYYYRYRYYRRYY`

S1 : `RYYRYYYYRYYYRYRYYRRYR`

To describe an arbitrary site in an ancestral sequence, we simply specify the probabilities that site might be occupied by either an `R` or `Y`. For instance the two numbers

$$(\mathcal{P}(S0 = \texttt{R}), \mathcal{P}(S0 = \texttt{Y})) = (p_R, p_Y) = (.5, .5)$$

would indicates an equal chance of each, while $(p_R, p_Y) = (.6, .4)$ would indicate a greater likelihood of a purine. Since our site must have either a purine or a pyrimidine, note the two probabilities must add to 1. If $(p_R, p_Y) = (.5, .5)$, then we can think of the ancestral base as being determined by the flip of a fair coin. If $(p_R, p_Y) = (.6, .4)$, then we can still think of the ancestral base as being determined by a coin flip, but the coin must be biased so that its '`R`' side lands up in 60% of a large number of tosses.

Although our model is describing only one site in the sequence, we view the data sequences as being many different trials of the same probabilistic process. Thus (p_R, p_Y), the probabilities that a site in an idealized ancestral sequence of infinite length is occupied by an `R` or `Y`, can be estimated by the frequencies at which these occur in the observed sequence. For example, the 21-base sequence S0 above has 9 `R`s and 12 `Y`s, which leads us to estimate $p_R = 9/21$ and $p_Y = 12/21$.

To justify this estimate, we are making what is often called an *i.i.d. assumption*, that each site behaves *independently* with an *identical distribution*. If the sites are independent and identically distributed, we may use the data for them all to infer things about the common process they all undergo.

With the ancestral sequence described, we now focus on probabilities of base substitutions as S0 evolves into S1. There are 4 possibilities here, $\texttt{R} \to \texttt{R}$, $\texttt{R} \to \texttt{Y}$, $\texttt{Y} \to \texttt{R}$ and $\texttt{Y} \to \texttt{Y}$, which in our data occurred in a total of 7, 2, 1, and 11 sites, respectively. It is most convenient to summarize this through conditional probabilities. For instance, the conditional probability that an ancestral `R` remains an `R` is denoted by

$$\mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{R}),$$

which we read as 'the probability that we have an `R` in S1 *given that* we had an `R` in S0.' For our data, using the i.i.d. assumption, we estimate

$$\mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{R}) = 7/9,$$

since of the 9 ancestral Rs, 7 remained Rs. Similarly we estimate

$$\begin{aligned}\mathcal{P}(\text{S1} = \texttt{Y} \mid \text{S0} = \texttt{R}) &= 2/9,\\ \mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{Y}) &= 1/12,\\ \mathcal{P}(\text{S1} = \texttt{Y} \mid \text{S0} = \texttt{Y}) &= 11/12.\end{aligned}$$

Notice

$$\begin{aligned}\mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{R}) + \mathcal{P}(\text{S1} = \texttt{Y} \mid \text{S0} = \texttt{R}) &= 1,\\ \mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{Y}) + \mathcal{P}(\text{S1} = \texttt{Y} \mid \text{S0} = \texttt{Y}) &= 1,\end{aligned}$$

as the total conditional probability of either an R or Y appearing in S1, assuming that the base in S0 is given, must be 1.

Now our model is summarized by six probabilities, which we organize into a row vector and a matrix:

$$\mathbf{p}_0 = \begin{pmatrix} p_R & p_Y \end{pmatrix} = \begin{pmatrix} 9/21 & 12/21 \end{pmatrix},$$

$$\begin{aligned} M &= \begin{pmatrix} \mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{R}) & \mathcal{P}(\text{S1} = \texttt{Y} \mid \text{S0} = \texttt{R}) \\ \mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{Y}) & \mathcal{P}(\text{S1} = \texttt{Y} \mid \text{S0} = \texttt{Y}) \end{pmatrix} \\ &= \begin{pmatrix} p_{RR} & p_{RY} \\ p_{YR} & p_{YY} \end{pmatrix} = \begin{pmatrix} 7/9 & 2/9 \\ 1/12 & 11/12 \end{pmatrix}. \end{aligned}$$

(Note the switch in order here for the state indications on the matrix entries, where $\mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{Y}) = p_{YR}$, for instance.) The rows of the matrix refer to ancestral states, while the columns refer to descendant ones. As a result, the entries across any row add to 1, though the column sums have no special value.

One point of this matrix notation is that the product $\mathbf{p}_0 M$ has meaningful entries:

$$\mathbf{p}_0 M = \begin{pmatrix} p_R & p_Y \end{pmatrix} \begin{pmatrix} p_{RR} & p_{RY} \\ p_{YR} & p_{YY} \end{pmatrix} = \begin{pmatrix} p_R p_{RR} + p_Y p_{YR} & p_R p_{RY} + p_Y p_{YY} \end{pmatrix}$$

where, for instance, the left entry is

$$\begin{aligned} p_R p_{RR} + p_Y p_{YR} &= \mathcal{P}(\text{S1} = R \mid \text{S0} = R)\mathcal{P}(\text{S0} = R) && (6.1)\\ &\quad + \mathcal{P}(\text{S1} = \texttt{R} \mid \text{S0} = \texttt{Y})\mathcal{P}(\text{S0} = \texttt{Y}) && (6.2)\\ &= \mathcal{P}(\text{S1} = \texttt{R}). && (6.3)\end{aligned}$$

Although this last equality follows from various formal multiplication and sum rules of probabilities, it can also be understood intuitively: The term on the right side of (6.1) gives the probability that we have an ancestral R that then remains an R. The term in (6.2) gives the probability that we have an ancestral Y that then changes to an R. Since these are the only ways we could have an R in the descendant site, by adding these, we are computing the probability that the descendant has an R, as claimed in equation (6.3).

Similarly, the right entry of the product $\mathbf{p}_0 M$ is $\mathcal{P}(\text{S1} = \texttt{Y})$. Thus if $\mathbf{p}_1$ denotes the probability distribution of `R`s and `Y`s for S1, we have shown

$$\mathbf{p}_1 = \mathbf{p}_0 M.$$

The matrix M is therefore not just a table to encapsulate the various probabilities of changes in the substitution process between the ancestral and descendant sequences; the multiplication of a vector by M actually 'does' the process.

What, then, might happen if the sequence S1 continues to evolve? Assuming circumstances are similar to those during the evolution of S0 to S1, and the elapsed time is similar, it's reasonable to hypothesize that we would obtain a sequence S2 whose `R`/`Y` composition would be described by

$$\mathbf{p}_2 = \mathbf{p}_1 M = \mathbf{p}_0 M^2.$$

Thinking of M as describing one time-step, we now have a *model* of sequence evolution using discrete time, with a descendant sequence after n time steps being described by

$$\mathbf{p}_n = \mathbf{p}_0 M^n.$$

The *parameters* of our model are a vector $\mathbf{p}_0$ of non-negative numbers that sum to 1, which we call the *root distribution vector*, and a matrix M of non-negative numbers whose rows sum to one, which we call a *Markov matrix* (or *stochastic matrix*). The root distribution describes the models starting conditions (the ancestral sequence) while the Markov matrix describes the substitution process in a single time step.

A continuous-time version

In the presentation above, we initially thought of M as describing evolution along a full edge of a tree. Then we shifted to imagining it as describing evolution for just one time-step, and that an edge might be many time steps long. While both of these views are useful in some settings, they essentially use a discrete notion of time.

While mutations do occur at discrete times, corresponding to generations of the evolving organism, this is not always the simplest way to view things. Since the span of a generation is usually quite small relative to evolutionary time scales, it is more common to describe the evolutionary process using a continuous notion of time.

In this formulation, we imagine that there are certain *rates* at which the various types of substitutions occur, and we organize them into a matrix such as

$$Q = \begin{pmatrix} q_{RR} & q_{RY} \\ q_{YR} & q_{YY} \end{pmatrix}.$$

Here, for instance q_{RY} denotes the instantaneous rate at which `R`s are replaced by `Y`s, and would be measured in units like (substitutions at a site)/(unit of time). Moreover $q_{RY} \geq 0$, or more likely $q_{RY} > 0$.

At first it seems nonsensical to say entry q_{RR} should be the rate at which Rs are replaced by Rs. But if some Rs are being replaced by Ys, then some Rs will cease to be Rs. Thus $q_{RR} \leq 0$ gives this rate of loss of Rs. Moreover, we must have $q_{RR} + q_{RY} = 0$ since these two rates must balance.

Considering the second row of Q similarly, we see that the entries in each row of Q must add to give 0 (unlike the 1 of the discrete-time Markov matrix). The off-diagonal entries of Q giving rate of changes from one state to another must be non-negative, so the diagonal entries must be non-positive.

Letting $\mathbf{p}_t = (p_R(t)\ p_Y(t))$ denote the distribution vector of purines and pyrimidines at time t, with $t = 0$ for the ancestral sequence, we have a system of differential equations:

$$\begin{aligned}\frac{d}{dt}p_R(t) &= p_R(t)q_{RR} + p_Y(t)q_{YR}\\ \frac{d}{dt}p_Y(t) &= p_R(t)q_{RY} + p_Y(t)q_{YY}.\end{aligned}$$

Expressing this system in matrix form, by letting $\mathbf{p}_t = (p_R(t), p_Y(t))$, yields

$$\frac{d}{dt}\mathbf{p}_t = \mathbf{p}_t Q. \tag{6.4}$$

(While systems of differential equations such as this are commonly covered in ordinary differential equations courses, they are usually presented using column vectors, with the matrix appearing to the left of the column vector. Our notation is more standard for probabilistic models, and is equivalent, through a matrix transpose.)

Before we sketch the solution of this system of differential equations, for motivation recall a similar differential equation that involves no vectors or matrices:

$$p'(t) = rp(t),\ p(0) = p_0.$$

This equation is often used to model a population, whose initial size is p_0, and which grows at a rate proportional to its size, with r being the constant of proportionality. The solution of this is

$$p(t) = p_0 e^{rt}.$$

Moreover, the exponential function appearing in this solution can be understood by a Taylor series:

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \ldots.$$

The system of differential equations (6.4) is solved similarly, using the initial values given by $\mathbf{p}_0$, with solution

$$\mathbf{p}_t = \mathbf{p}_0 e^{Qt}.$$

This formula involves the exponential of a matrix $A = Qt$, which can be defined by the usual Taylor series formula, where all terms are reinterpreted as matrices. For any square matrix A

$$e^A = I + A + \frac{1}{2}A^2 + \frac{1}{3!}A^3 + \frac{1}{4!}A^4 + \dots. \tag{6.5}$$

Note that this is *not* the same as applying the usual exponential function to the entries of A individually; the powers that appear in the Taylor series result in interaction between the various entries of A.

While the Taylor series for the matrix exponential can provide a good conceptual understanding, to compute matrix exponentials easily requires more theory from linear algebra. Provided A can be diagonalized, then writing $A = S\Lambda S^{-1}$ with Λ the diagonal matrix of eigenvalues of A and S the matrix whose columns are the corresponding (right) eigenvectors, we also have (Exercise 16)

$$e^A = Se^{\Lambda}S^{-1}, \tag{6.6}$$

where e^{Λ} is a diagonal matrix with diagonal entries the exponentials of the entries of Λ.

Returning to our model, since $\mathbf{p}_t = \mathbf{p}_0 e^{Qt}$, it is natural to define

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

as the Markov matrix which encodes the substitutions of bases that occur when an amount of time t passes. The entries of $M(t)$ are thus the conditional probabilities of the various substitutions being observed as we compare sequences from time 0 to those from time t, so that $M(t)$ is a single matrix describing the mutation along an edge representing a time of length t.

Our formulation of the model through the rate matrix Q has made the assumption that over the full time interval state changes have been occurring at exactly the same rates. Although we can be a little less strict than this by imagining our clock runs slow or fast at different times, we are still assuming the rates at which Rs becomes Ys and the rate at which Ys become Rs are in fixed proportion to each other. Thus the state substitution process is viewed as uniform, except perhaps for clock speed changes

Note the important distinctions between a Markov matrix M and a rate matrix Q. The entries of M are probabilities, and hence lie between 0 and 1, its rows sum to 1, and it describes a substitution process either along an entire edge, or over a discrete time step. The entries of Q are not probabilities, but rather rates describing the instantaneous substitution process, and hence do not need to be between 0 and 1. The off-diagonal entries, however, cannot be negative, and the ones on the diagonal cannot be positive. The rows of Q sum to 0 . Applying the matrix exponential function to the product of a rate matrix and an elapsed time gives a Markov matrix.

6.2 Markov Models on Trees

We now more carefully define a rather general model of DNA evolution along a tree. Although we use 4-state characters, the connection to the 2-state character version developed in the previous section should be clear. It is also straightforward to formulate a 20-state version appropriate for describing protein sequences, or a 61-state version for a codon model ($61 = 4^3 - 3$, since the 3 'stop' codons are generally excluded).

A general DNA base-substitution model

The possible states of our character are `A`,`G`,`C`,`T`. We always use this order for the four bases, so that the purines precede pyrimidines, with each in alphabetical order.

Consider a fixed rooted tree T^ρ. Then parameters for the *general Markov model* model on T^ρ consist of the following:

1) A *root distribution vector* $\mathbf{p}_\rho = (p_A, p_G, p_C, p_T)$, with all entries non-negative and $p_A + p_G + p_C + p_T = 1$. We interpret these entries as giving the probabilities that an arbitrary site in a DNA sequence at ρ is occupied by the corresponding base, or, equivalently, as the frequencies with which we would expect to observe these bases in a sequence at ρ. (Note the i.i.d. assumption here.)

2) For each edge $e = (u, v)$ of T directed away from ρ, a 4×4 *Markov matrix*, M_e, whose entries are non-negative and whose rows sum to 1. The i, j-entry of M_e we interpret as the conditional probability that if base i occurs at the site in the parent vertex on the edge, then base j occurs at the descendant vertex. Using abbreviated notation with $p_{ij} = \mathcal{P}(S_v = j \mid S_u = i)$, where S_u, S_v denote sequences at u and v respectively, we let
$$M_e = \begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AT} \\ p_{GA} & p_{GG} & p_{GC} & p_{GT} \\ p_{CA} & p_{CG} & p_{CC} & p_{CT} \\ p_{TA} & p_{TG} & p_{TC} & p_{TT} \end{pmatrix}.$$

Note that the only mutations described by this model are base substitutions. No events such as insertions, deletions, or inversions are included in the formulation.

This basic model, called the *general Markov model*, will be our starting point. Shortly we will discuss more restrictive models, where we place additional requirements on the model parameters, and then later will present some generalizations. Although the general Markov model is seldom directly used for data analysis, it provides the basic framework for all models in widespread use.

As a final notational point, since we have fixed the ordering `A`, `G`, `C`, `T`, and will often need to refer to particular entries of vectors and matrices, it will also

be convenient to sometimes use numbers to refer to the bases, with

$$1 = \texttt{A},\ 2 = \texttt{G},\ 3 = \texttt{C},\ 4 = \texttt{T}.$$

A common rate-matrix model

Rather than allow completely unrelated Markov matrices for the base substitution process on each edge of the tree, most model-based phylogenetic analyses specify that the substitution processes along the various edges of the tree have some commonality. The usual way to do this is to replace point (2) above with the following:

2a) A continuous-time 4×4 rate matrix Q, whose rows add to 0, and whose off-diagonal entries are nonnegative. With

$$Q = \begin{pmatrix} q_{AA} & q_{AG} & q_{AC} & q_{AT} \\ q_{GA} & q_{GG} & q_{GC} & q_{GT} \\ q_{CA} & q_{CG} & q_{CC} & q_{CT} \\ q_{TA} & q_{TG} & q_{TC} & q_{TT} \end{pmatrix},$$

we interpret an entry q_{ij} as the instantaneous rate (in substitutions at a site per unit time) at which base i is replaced by base j.

2b) For each edge e of the tree a non-negative scalar length t_e. Then the Markov matrix

$$M_e = M(t_e) = e^{Qt_e}$$

can be interpreted as above in (2) for the edge e.

One attractive feature of this model is that it gives meaning to edge lengths t_e in a metric tree as specifying the 'amount' of substitution that will occur along that edge. With the general Markov model it is not immediately clear how to associate a single number to an edge to indicate such a measure. On the other hand, the assumption of a common process across the entire tree is a strong one (though overwhelmingly common in data analysis). In practice, the choice of Q is usually further restricted, as we will discuss later.

The distribution of character states at the leaves

With either general Markov parameters, or continuous-time parameters for a model on a tree T^ρ specified, we can compute probabilities of observing any particular combination of bases in aligned sequences at the vertices of a tree. We focus only on the probabilities of various base observations at the leaves, since that is the only sort of data that we typically can obtain from organisms. (For statistical inference, we will *not* know appropriate parameters to use for our model, but will have to somehow infer them from leaf data, inverting the process we are explaining here. That is the subject of later chapters.)

To see how to compute leaf distributions, we begin with several examples for very simple trees.

A one-edge tree

This case is similar to that discussed in the purine/pyrimidine example which began the chapter. But now we adopt the viewpoint that we know the root distribution, and a Markov matrix describing the substitution process over the edge, and we wish to compute the probability of observing each base at the single leaf of the tree.

If $\mathbf{p}_\rho$ and M_e are the parameters on a one edge tree from root ρ to descendant S1, then

$$\mathbf{p}_1 = \mathbf{p}_\rho M_e$$

gives a vector of probabilities of observing the four bases at any site in the descendant sequence. For instance, from the definition of matrix multiplication, the first entry of $\mathbf{p}_1$, which should refer to the probability of observing an `A` in the descendant sequence, is

$$p_1 M_e(1,1) + p_2 M_e(2,1) + p_3 M_e(3,1) + p_4 M_e(4,1) = \\ p_A p_{AA} + p_G p_{GA} + p_C p_{CA} + p_T p_{TA},$$

which has the claimed probabilistic interpretation. We are summing 4 terms for the 4 possible ancestral bases; each term gives the probability we had that ancestral base and it changed to (or remained as) an `A`. The other entries of $\mathbf{p}_1$ are produced similarly.

A two-edge tree

Suppose now S1, S2 are two children of a root ρ, with M_i the Markov matrix on the edge leading to Si, as in Figure 6.1.

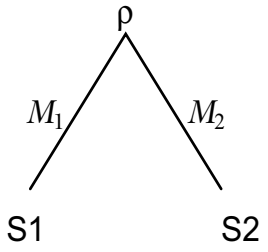

Figure 6.1: A two-edge tree

Then, if we ignore S2, we can compute the probabilities of the various bases appearing in S1 by the product $\mathbf{p}_\rho M_1$, and similarly for S2. But what is the *joint probability* that at a particular site we observe base i in S1, and base j at S2? Since we will have to consider every pair i, j of bases at the two leaves, we should organize the probabilities of these observations into a matrix P, with $P(i,j)$ denoting the probability of observing i at S1 and j at S2.

For an observation (i, j) to be produced, we might have any base k at ρ. Then this k must become an i in S1, and a j in S2. For a particular k, this means the probability is

$$p_k M_1(k, i) M_2(k, j).$$

But since k could be anything, we need to sum over the possibilities to get

$$P(i, j) = \sum_{k=1}^{4} p_k M_1(k, i) M_2(k, j) = p_1 M_1(1, i) M_2(1, j) + p_2 M_1(2, i) M_2(2, j) + p_3 M_1(3, i) M_2(3, j) + p_4 M_1(4, i) M_2(4, j).$$

This can be more succinctly expressed as a matrix equation. Letting $\text{diag}(\mathbf{v})$ denote a diagonal matrix with the entries of a vector $\mathbf{v}$ on its diagonal, the 4×4 matrix P whose entries give the joint distribution of bases at the leaves is given by

$$P = M_1^T \, \text{diag}(\mathbf{p}_\rho) M_2. \tag{6.7}$$

We leave checking this claim as Exercise 7.

We should also point out that once all the entries of any joint distribution P are computed, it is easy to recover the distribution of bases at a single leaf, by summing over the other indices, a process usually called *marginalization*. For instance, the probabilities of observing the base j at S2, without regard to what appears at S1 is found by summing over the index corresponding to S1, giving

$$\sum_{i=1}^{4} P(i, j).$$

Checking that this agrees with the jth entry of $\mathbf{p}_\rho M_2$ is Exercise 8.

A many-edge tree

Suppose now T^ρ has many edges, such as in Figure 6.2

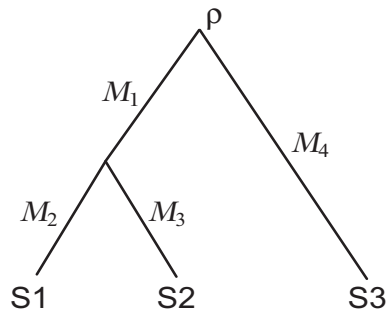

Figure 6.2: A small, many-edged tree

In general, we will not be able to give a simple formula for the joint distribution of bases at the leaves of the tree using standard matrix notation. Indeed, we should not expect to, because the joint distribution itself will be specified by an array of more than 2 dimensions. For instance, in the tree above, since there are 3 leaves, the joint distribution tensor P will be a $4 \times 4 \times 4$ array, with the entry $P(i, j, k)$ giving the probabilities that a site has bases i, j, k at the leaves $S1, S2, S3$, respectively. Using slightly different terminology, $P(i, j, k)$ gives the probability of observing the *pattern* ijk at a site in aligned sequences for $S1, S2, S3$.

A formula for any particular entry $P(i, j, k)$ of the joint distribution tensor is easily given, however, by following the same sort of reasoning as for the two-edge tree. We simply imagine each possible assignments of bases to all internal nodes of the tree, and then write a product expression for the probability of this particular evolutionary history. Afterwards, we sum over all such interior node assignments, since these are all the distinct ways of obtaining the pattern we are interested in. For example, for the tree above we have

$$P(3, 1, 1) = \sum_{i=1}^{4} \sum_{j=1}^{4} p_i M_1(i, j) M_2(j, 3) M_3(j, 1) M_4(i, 1) \tag{6.8}$$

as one of the 64 entries of P. Since there are 2 interior nodes in this tree, this sum has $4^2 = 16$ terms in it. Since there are 4 edges in the tree, each term is a product of $1 + 4 = 5$ parameters, one for each edge and one for the root. The other 63 entries are given by quite similar formulas.

For a tree relating more taxa, of course the formulas get more complicated, but the idea of how they are formed should be clear. It is worth emphasizing, though, that the formulas for the entries of P depend not just on the number of taxa, but also on the particular tree topology. Indeed, until we fix a tree topology to consider, we can't even relate Markov matrices to the particular edges.

Assumptions of the models

While we've been calling the matrices of conditional probabilities Markov matrices, we should say a little more about this terminology.

A probabilistic model of a system is called a *Markov model* if it incorporates an assumption of the following sort:

> The behavior of the system over any given time period depends only on the state the system is in at the beginning of that period. The earlier history of the system can affect what happens in the future only through having already affected the system's current state.

More formally, in a Markov model the probability that a particular state change occurs given the system is in state i is the same as the probability of the same change, given any entire earlier history of states ending in state i. In particular, a 'memory' of what state changes occurred during earlier times is

useless for predicting future changes. We say the probabilities of state changes are *independent* of the earlier history.

The model we have given here, of molecular evolution occurring through random base substitutions, satisfies the Markov assumption. Since time proceeds from the root to the leaves, and the probabilities of the various possible state changes on any given edge depend only on the state at the ancestral node on that edge, the above condition is met.

In fact, the general Markov model makes an even stronger, but equally reasonable, assumption that mutation process on one edge is not affected by what occurs on any edge that is not ancestral to it. Any correlations we observe between state observations at two leaves arise solely from the state of their most recent common ancestor.

In applications of the model we will also assume that each site in the sequence behaves identically, and independently of every other site (i.i.d). Though the model describes substitutions at a single site, it applies equally well to every site in the sequence. We used this assumption in our introductory example in order to find the various probabilities we needed from our sequence data, by thinking of each site as an independent trial of the same probabilistic process. We will continue to do this when we use more elaborate methods to determine appropriate parameter values from data.

The i.i.d. assumption is probably *not* very reasonable for DNA in many circumstances. For instance, since the genetic code allows for many changes in the third site of each codon that have no affect on the product of the gene, one could argue that substitutions in the third sites might be more likely than in the first two sites, violating the assumption that each site behaves identically. Also, since genes may lead to the production of proteins which have functional roles in life's processes, the chance of change at one site may well be tied to changes at another, through the need for the gene product to have a particular form. This violates the assumption of independence. A similar issue arises from the *secondary structure* of RNA formed by the transcription of genes. In the single-stranded RNA sequence, some bases form bonds to bases further down the sequence, producing features called 'stems' and 'loops' . This means that mutations at one site may well be tied to mutations at sites that are not even nearby in the sequence.

It's easy to come up with a host of other problems with the i.i.d. assumption: If sequences encode both coding and noncoding regions, should they really evolve the same way? Might some genes tend to mutate faster than others? Might not the bases in some sites be so crucial to a sequence's functioning that they are effectively prevented from mutating? Modifications to the basic modeling framework will at least partially address some of these issues, but some form of an i.i.d. assumption is always needed in order to have well-behaved inference in standard statistical frameworks. What matters is not that it is exactly true, but rather that it is a good enough approximation of the truth to apply statistical methods.

A useful analogy for understanding why the i.i.d. assumption is so crucial,

is to imagine data from a hundred coin flips. If your model is that the same (or an essentially identical coin) was flipped in the same way a hundred times, then the i.i.d. assumption holds, and by computing the frequency of heads, you can obtain a good estimate of the probability a single coin flip gives heads.

However, if you instead insist that these coin flips should be modeled as a hundred flips of varying coins, so we drop the identically distributed assumption, it is possible that all the heads were produced by coins weighted to always land heads up, and the tails by ones biased oppositely. If this were the case, then computing the frequency of heads among the hundred would not tell us anything useful about a single coin flip, or enable us to say anything about what might happen at additional sites if the sequences were longer.

If we instead drop the independent assumption, then we might flip one fair coin, and have it determine all the other outcomes, so that our data could only be all heads, or all tails. In this case, the frequency of heads among the hundred would either be 0 or 1, but that again tells us nothing about the probability of outcomes of a single coin flip.

6.3 Jukes-Cantor and Kimura Models

The general Markov model, or its continuous-time variant, is usually further restricted to special forms for data analysis. We present a few of these, starting from the most restricted, and then relaxing assumptions.

The Jukes-Cantor model

The simplest Markov model of base substitution, the *Jukes-Cantor* model, adds several additional assumptions to the general Markov model.

First, it assumes the root distribution vector describes all bases occurring with equal probability in the ancestral sequence. Thus

$$\mathbf{p}_\rho = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}.$$

Second, as a continuous-time model it assumes a rate matrix of the form

$$Q = \begin{pmatrix} -\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & -\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & -\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & -\alpha \end{pmatrix}. \tag{6.9}$$

This indicates that the rate of all specific base changes, `A` $\leftrightarrow$ `T`, `A` $\leftrightarrow$ `C`, `A` $\leftrightarrow$ `G`, `C` $\leftrightarrow$ `T`, `C` $\leftrightarrow$ `G`, and `T` $\leftrightarrow$ `G` are the same, $\alpha/3$. The total rate at which any specific base is changing to the other 3 bases is therefore α.

The associated Markov matrix on an edge of length t can now be calculated as

$$M(t) = e^{Qt}.$$

This requires first finding eigenvectors and eigenvalues for Q (see exercise 18) to obtain the diagonalization formula

$$Q = S\Lambda S^{-1},$$

with

$$S = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad \Lambda = \operatorname{diag}\left(0, -\frac{4}{3}\alpha, -\frac{4}{3}\alpha, -\frac{4}{3}\alpha\right). \tag{6.10}$$

Thus the Jukes-Cantor Markov matrix for an edge of length t is

$$\begin{aligned} M(t) &= e^{Qt} \\ &= Se^{\Lambda t}S^{-1} \\ &= S^{-1}\operatorname{diag}\left(1, e^{-\frac{4}{3}\alpha t}, e^{-\frac{4}{3}\alpha t}, e^{-\frac{4}{3}\alpha t}\right) S \\ &= \begin{pmatrix} 1-a & a/3 & a/3 & a/3 \\ a/3 & 1-a & a/3 & a/3 \\ a/3 & a/3 & 1-a & a/3 \\ a/3 & a/3 & a/3 & 1-a \end{pmatrix}, \end{aligned}$$

where

$$a = a(t) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}\alpha t}\right). \tag{6.11}$$

Since the entries of $M(t)$ are probabilities, we interpret $a(t)$ as the probability that any specific base at time 0 will have changed to any of the 3 other bases at time t. This might have happened by only one base substitution occurring, or it might have happened through a succession of substitutions. The continuous-time model accounts for all possible ways the final state could have been achieved from the initial one. We are likely, of course, to use a different value of a for each edge of the tree, since the formula for a depends on the edge length. Larger values of t produce larger values of a, as more mutation occurs on that edge.

The Jukes-Cantor model also implies a *stable base distribution* at all vertices of the tree. To see this, we simply compute the effect of the substitution process as we proceed down an edge:

$$\begin{aligned} \mathbf{p}_\rho M &= \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix} \begin{pmatrix} 1-a & a/3 & a/3 & a/3 \\ a/3 & 1-a & a/3 & a/3 \\ a/3 & a/3 & 1-a & a/3 \\ a/3 & a/3 & a/3 & 1-a \end{pmatrix} \\ &= \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}. \end{aligned}$$

Of course the simple, highly-symmetric form of the Jukes-Cantor matrix means that for every substitution from state i to j we should expect a substitution from j to i, so that is it not surprising the base distribution never changes.

We will return to equation (6.11) in the next chapter, as αt has an important interpretation. Since α is a mutation rate, measured in units (number of substitutions at a site)/(unit of time), and t represents an amount of time, the product tells us the number of substitutions that should occur at a site over the elapsed time t. While both α and t depend on our choice of units of time, the product has a meaning independent of those units.

Mutation rates such as α for DNA in real organisms are not easily found, since estimating them appears to require both an ancestral and descendant sequence, and knowledge of the evolutionary time separating them. For reasons that will be explained in the next chapter, its possible to avoid finding an ancestral sequence, but we do need an independent estimate of the divergence time of two descendants from their common ancestor, perhaps obtained from a fossil. Various estimates of α place it around 1.1×10^{-9} mutations per site per year for certain sections of chloroplast DNA of maize and barley and around 10^{-8} mutations per site per year for mitochondrial DNA in mammals. The mutation rate for the influenza A virus has been estimated to be as high as .01 mutations per site per year. The rate of mutation is generally found to be a bit lower in coding regions of nuclear DNA than in non-coding DNA.

The Kimura models

The Jukes-Cantor model is a particularly simple model of mutation since it depends on only one single parameter α to specify the rate of mutation.

The model can be made more flexible by allowing several parameters. A good example of this is the *Kimura 2-parameter* model, which allows for different probabilities of transitions and transversions.. If we let

$$Q = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix},$$

then β is the rate at which each state undergoes transitions, and 2γ is the rate that any state undergoes transversions, with transversions equally likely to produce either possible outcome. The entries denoted by * should be $-\beta - 2\gamma$, since that is what is needed so that row sums are 0.

Although we leave the details as an exercise, a computation of the matrix exponential shows the associated Markov matrix has the form

$$M(t) = e^{Qt} = \begin{pmatrix} * & b & c & c \\ b & * & c & c \\ c & c & * & b \\ c & c & b & * \end{pmatrix},$$

with

$$b = \frac{1}{4}(1 - 2e^{-2(\beta+\gamma)t} + e^{-4\gamma t}), \quad c = \frac{1}{4}(1 - e^{-4\gamma t}).$$

Here the entries denoted * are $1-b-2c$, since the rows must add to 1. Notice that if the probabilities of a transition and each transversion are equal so $\beta = \gamma$, then this model includes the Jukes-Cantor one as a special case with $\alpha = 3\beta = 3\gamma$.

An even more general model is the Kimura 3-parameter model, which assumes a rate matrix of the form

$$Q = \begin{pmatrix} * & \beta & \gamma & \delta \\ \beta & * & \delta & \gamma \\ \gamma & \delta & * & \beta \\ \delta & \gamma & \beta & * \end{pmatrix},$$

which leads to a Markov matrix of the form

$$M = \begin{pmatrix} * & b & c & d \\ b & * & d & c \\ c & d & * & b \\ d & c & b & * \end{pmatrix}.$$

By an appropriate choice of the parameters, this includes both the Jukes-Cantor and Kimura 2-parameter models as special cases. (While the structure of the Kimura 2-parameter model is suggested by biology, the generalization to the 3-parameter is primarily motivated by mathematical properties.)

Both Kimura models also assume that the root distribution vector is the uniform one,

$$\mathbf{p}_\rho = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}.$$

A quick calculation shows that for either of these models this distribution will be stable, occurring at all vertices in the tree.

There is a clear relationship between the form of the rate matrix for the Jukes-Cantor and Kimura models and the form of the associated Markov matrices, with both exhibiting a very similar pattern of entries. However, this is a very special feature of these models, and such simple relationships do not occur more generally. While the entries of the Markov matrix will always be expressible in terms of those of the rate matrix and t, there is generally no easy way to find these expressions, except as outlined in the computation above for the Jukes-Cantor model.

A note of caution: When a Kimura model is used on a tree, there are two ways it might be used. It may be that one Kimura rate matrix (with fixed β,γ) is used for the entire tree with the 1 parameter of a scalar length assigned to each edge. Alternately, one may assign different and unrelated Kimura Markov matrices on each edge (giving 2 or 3 parameters per edge). The first approach is more common in standard approaches to phylogenetic inference, and is likely to be what is assumed in software performing Maximum Likelihood or Bayesian analyses. The second is a more general model, and often appears in more theoretical works, especially those dealing with Hadamard conjugation.

6.4 Time-reversible Models

An important feature of the Jukes-Cantor and Kimura models is that they are *time-reversible*. What this means is that given an ancestral and a descendant sequence separated by one edge of the tree, if we reverse the flow of time, interchanging which sequence we view as ancestor and descendant, we would describe evolution by exactly the same model parameters.

To see what this means mathematically, suppose $\mathbf{p}$ is the ancestral base distribution and M the Markov matrix describing the descent. Let P denote the joint distribution of bases in the ancestral and descendant sequences, so $P(i,j)$ gives the probability of an ancestral i and a descendant j appearing at a site. Then since $P(i,j) = p_i M(i,j)$, we have the matrix equation

$$P = \operatorname{diag}(\mathbf{p})M.$$

Now interchanging our view of what sequence is ancestral means interchanging the order of indices, or equivalently transposing P. Thus time reversibility means $P = P^T$, or

$$\operatorname{diag}(\mathbf{p})M = (\operatorname{diag}(\mathbf{p})M)^T = M^T \operatorname{diag}(\mathbf{p}). \tag{6.12}$$

In fact, this equation implies the intuitive fact that time-reversibility requires that $\mathbf{p}$ be a stable base distribution for M. (See Exercise 27.)

For the Jukes-Cantor and Kimura models it is easily checked that equation (6.12) holds in the discrete-time formulation. However there are many other possible parameter choices for a time-reversible model.

To elaborate, in the continuous-time formulation of a model, for time-reversibility to hold we need

$$\operatorname{diag}(\mathbf{p})Q = Q^T \operatorname{diag}(\mathbf{p}). \tag{6.13}$$

If we let

$$\mathbf{p} = \begin{pmatrix} p_A & p_G & p_C & p_T \end{pmatrix}$$

be the root distribution, then a little work (Exercise 28) shows Q should be of the form

$$Q = \begin{pmatrix} * & p_G\alpha & p_C\beta & p_T\gamma \\ p_A\alpha & * & p_C\delta & p_T\epsilon \\ p_A\beta & p_G\delta & * & p_T\eta \\ p_A\gamma & p_G\epsilon & p_C\eta & * \end{pmatrix}, \tag{6.14}$$

with $\alpha, \beta, \gamma, \delta, \epsilon, \eta \geq 0$ and where the diagonal entries are chosen so rows sum to zero.

Perhaps the most commonly-used basic continuous-time model used in data analysis is the *general time-reversible model* (GTR). It is specified by the following parameters:

1. an arbitrary choice of a root distribution $\mathbf{p} = \begin{pmatrix} p_A & p_G & p_C & p_T \end{pmatrix}$
2. arbitrary choices of 6 parameters $\alpha, \beta, \gamma, \delta, \epsilon, \eta$, with the common time-reversible rate matrix Q on all edges, given by equation 6.14, and

3. lengths of edges in the tree

The form of Q then ensures that $\mathbf{p}$ is stable under the model, and thus is the distribution of bases at all vertices in the tree.

It's easy to see that with special choices of the parameters, the GTR includes the Jukes-Cantor and Kimura models. But it has more flexibility due to the larger number of parameters, and thus is capable of describing a larger class of data sets well.

Practically, the GTR model seems to be a good compromise between simplicity and complexity. Having a common rate matrix reduces the number of parameters considerably from what would be needed without this assumption. This can help avoid 'overfitting' of data, keeping the variance in the inferred tree lower with the same amount of data. It also allows for faster run-times of software, as there are fewer parameters to be varied in searching for a good fit. A common rate matrix also imposes some commonality on the mutation process throughout the tree, which in some circumstances is biologically reasonable. That the base distribution is stable may also be reasonable, if all the data sequences have a similar base composition. Time reversibility means we will be able to ignore issues of root location when we try to find optimal trees, thereby reducing search time slightly. Nonetheless, it's hard to imagine a full justification of this model solely on biological grounds. Base compositions are not always identical across taxa, and the model cannot capture that. It is also hard to imagine a biological justification for time-reversibility.

While models that drop some of the characteristics of the GTR model are occasionally used in data analysis, the vast majority of phylogenetic analyses currently use the GTR model, or special cases of it, as their basis. While many of these special cases have their own names, they can all be viewed as simply imposing relationships among the GTR parameters to reduce their number. For instance the F81 model (introduced by Felsenstein in a paper in 1981) allows for any choice of the base distribution, but sets the remaining GTR parameters $\alpha = \beta = \gamma = \delta = \epsilon = \eta = 1$. It can thus be viewed as an analog of the Jukes-Cantor model that does not require equidistribution of the bases. The HKY model (introduced by Hasegawa, Kishino, and Yano) also allows any base distribution, but sets $\beta = \gamma = \delta = \epsilon = 1$ and $\alpha = \eta = \kappa$ where the parameter κ can be viewed as a transition/transversion rate ratio. Thus it is an analog of the Kimura 2-parameter model for unequal base distribution.

While its tempting to think that the GTR model should always be the best one to use since it makes fewer special assumptions, that in fact is not the case. In general its desirable to use a model that fits the data reasonably well, but has few parameters. More restrictive models of this sort can lead to better inference, since they avoid overfitting data.

6.5 Exercises

1. Suppose ancestral and descendant sequences of purines and pyrimidines are

   S0 = RRYRYYRRRRYYYRYYRRYYRYYR
   S1 = RYYRYRRRRRRYYRRYRRYYRYYY

   Use this data to estimate a distribution vector for S0 and a Markov matrix describing the mutation process from S0 to S1.

2. The joint frequencies of purines and pyrimidines in ancestral and descendant sequences S0 and S1 is summarized in Table 6.1. Use this data to estimate a distribution vector for S0 and a Markov matrix describing the mutation process from S0 to S1.

$S0 \diagdown S1$	R	Y
R	183	32
Y	15	211

   Table 6.1: Frequencies from site comparisons for a pair of sequences

3. An ancestral DNA sequence of 40 bases was

   CTAGGCTTACGATTACGAGGATCCAAATGGCACCAATGCT,

   but in a descendant it had mutated to

   CTACGCTTACGACAACGAGGATCCGAATGGCACCATTGCT.

   a. Give an initial base distribution vector and a Markov matrix to describe the mutation process.

   b. These sequences were actually produced by a Jukes-Cantor simulation. Is that surprising? Explain. What value would you choose for the Jukes-Cantor parameter a to approximate your matrix by a Jukes-Cantor one?

4. Data from two comparisons of 400-base ancestral and descendant sequences are shown in Table 6.2.

   a. For one of these pairs of sequences a Jukes-Cantor model is appropriate. Which one, and why?

   b. What model would be appropriate for the other pair of sequences? Explain.

5. The Markov matrices that describe real DNA mutation tend to have their largest entries along the main diagonal in the (1,1), (2,2), (3,3), and (4,4) positions. Why should this be the case?

$S_0 \diagdown S_1$	A	G	C	T
A	92	15	2	2
G	13	84	4	4
C	0	1	77	16
T	4	2	14	70

$S_0' \diagdown S_1'$	A	G	C	T
A	90	3	3	2
G	3	79	8	2
C	2	4	96	5
T	5	1	3	94

Table 6.2: Frequencies from 400 site comparisons for two pairs of sequences

6. On the tree in Figure 6.2, consider the Jukes-Cantor model where all the M_i have $a = 0.1$. Compute the probabilities of observing each of the following characters

   a) S1: `G`, S2: `G`, S3: `G`

   b) S1: `G`, S2: `G`, S3: `T`

   c) S1: `G`, S2: `T`, S3: `G`

   Are the largest and smallest of these probabilities what you might have expected? Explain.

7. Check that the matrix equation (6.7) is correct.

8. Let u denote a column vector with all entries 1. Explain why for a Markov matrix M that $Mu = u$. Then use this fact to show that summing over the first index of

$$M_1^T \operatorname{diag}(\mathbf{p}_\rho) M_2$$

   gives

$$\mathbf{p}_\rho M_2.$$

   Interpret this as explaining why the marginalization of a joint distribution of bases at two leaves gives the distribution of bases at one leaf.

9. For the 4-taxon tree $((a, b), (c, d))$ give a formula like that in equation (6.8) for the joint distribution of bases at the leaves in terms of parameters of the general Markov model. Do the same for the 4-taxon tree $(((a, b), c), d)$.

10. Suppose T^ρ is a rooted binary n-taxon tree. How many terms would be summed in the formula analogous to Equation (6.8) for computing the probability of a particular character? How many parameters would be multiplied in each term of this sum?

11. Make up a 4×4 Markov matrix M with all positive entries, and an initial $\mathbf{p}_0$. To be biologically realistic, make sure the diagonal entries of M are the largest.

    a. Use a computer to observe that after many time steps $\mathbf{p}_t = \mathbf{p}_0 M^t$ appears to approach some equilibrium. Estimate the equilibrium vector as accurately as you can.

b. Is your estimate in part (a) a left eigenvector of M with eigenvalue 1? If not, does it appear to be close to having this property?

c. Use a computer to compute the eigenvectors and eigenvalues of M. (In MATLAB the command `[S D]=eig(M)` computes right eigenvectors, so you will have to apply it to `M'`). Is 1 an eigenvalue? Is your estimate of the equilibrium close to its eigenvector?

12. Express the Kimura 2-parameter model using a 4×4 matrix, but with the bases in the order `A`,`C`,`G`,`T`. How is your matrix different from the one presented in the text? Explain.

13. Suppose we wish to model molecular evolution not at the level of DNA sequences, but rather at the level of the proteins that genes encode.

    a. Create a simple one-parameter mathematical model (similar to the Jukes-Cantor model) describing the process. You will need to use that there are 20 different amino acids from which proteins are constructed in linear chains.

    b. In this situation, how many free parameters would the general Markov model have on a single edge of the tree?

14. Do the data in Exercise 1 appear to be fit by a time-reversible model? Explain.

15. Suppose you have compared two sequences S_α and S_β of length 1000 sites and obtained the data in Table 6.3 for the number of sites with each pair of bases.

$S_\alpha \diagdown S_\beta$	A	G	C	T
A	105	15	15	15
G	25	175	25	25
C	35	35	245	35
T	25	25	25	175

Table 6.3: Frequencies of $S_\alpha = i$ and $S_\beta = j$ in 1000 site sequence comparison

    a. By marginalizing, compute the base distribution for each of the taxa individually. Does it appear to be stable?

    b. Assuming S_α is the ancestral sequence, find an initial base distribution $\mathbf{p}_0$ and a Markov matrix M to describe the data. Is your matrix M Jukes-Cantor? Is $\mathbf{p}_0$ a stable distribution for M?

    c. Assuming S_β is the ancestral sequence, find an initial base distribution $\mathbf{p}'_0$ and a Markov matrix M' to describe the data. Is your matrix M' Jukes-Cantor? Is $\mathbf{p}'_0$ a stable distribution for M'?

    You should have found that one of your matrices was Jukes-Cantor and the other was not. This can't happen if both S_α and S_β have base distribution $(.25, .25, .25, .25)$.

16. Assuming $A = S\Lambda S^{-1}$, show that equation (6.5) implies equation (6.6). Also explain why if a matrix B is diagonal then e^B is also diagonal, with diagonal entries obtained by exponentiating those of B.

17. Show that if the rows of Q sum to 0, then the rows of e^{Qt} sum to 1. (Hint: Use the Taylor series for the exponential, and the fact that the rows of Q summing to 0 is expressible as $Q\mathbf{u} = 0$, where $\mathbf{u}$ is a column vector with all entries 1.)

18. Check that the eigenvectors and eigenvalues of the Jukes-Cantor rate matrix Q in equation (6.9) are those given in equations (6.10).

19. The matrix S of eigenvectors of a Jukes-Cantor matrix that is given in equation 6.10 is quite special, and is sometimes called a *Hadamard matrix*. Compute S^2, and explain why this shows that $S^{-1} = \frac{1}{4}S$.

20. The formula for e^{Qt} for the Jukes-Cantor model in equation (6.11) and its predecessor can be used to understand the effect of an infinite amount of mutation, by letting $t \to \infty$.

    a. If $\alpha > 0$, what is $\lim_{t\to\infty} e^{-\frac{4}{3}\alpha t}$.

    b. Use this to explain why

$$e^{Qt} \to \begin{pmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{pmatrix}.$$

    Note that each of the rows of this matrix is the stable distribution.

    Explain informally why this limit is what you should have expected.

    c. Why did we exclude $\alpha = 0$ from our analysis?

21. Based on the last problem, one might conjecture that powers of a Markov matrix all of whose entries are positive approach a matrix whose rows are the stable distribution. On a computer, investigate this experimentally by creating a Markov matrix, computing very high powers of it to see if the rows become approximately the same, and then checking whether this row is a left eigenvector with eigenvalue 1 of the original matrix.

22. Let $M(a)$ denote a Jukes-Cantor Markov matrix with parameter a.

    a) Show the product $M(a_1)M(a_2)$ is $M(a_3)$, and give a formula for a_3 in terms of a_1, a_2.

    b) If a Jukes-Cantor matrix $M(a)$ describes the evolution of one sequence to another, then the Hamming distance estimates a. Explain why the formula you found in part (a) indicates the Hamming distance is not usually additive in the sense defined in Exercise 6 of Chapter 5.

    c) Explain why the formula you found in (a) indicates the Hamming distance is approximately additive when its values are small.

23. Show the product of two Kimura 3-parameter Markov matrices is again a Kimura 3-parameter Markov matrix.

24. Show the Kimura 3-parameter matrices (both Markov and rate) have the same eigenvectors as those given in the text for the Jukes-Cantor matrices. What are the eigenvalues of the Kimura 3-parameter rate matrices?

25. Use the results of the last problem to find the entries of e^{Qt} where $Q = Q(\beta, \gamma, \delta)$ is the Kimura 3-parameter rate matrix. Your result should be a Kimura 3-parameter Markov matrix. Give formulas for the Markov matrix entries b, c, d in terms of β, γ, δ, t. Show that in the special case of the Jukes-Cantor and Kimura 2-parameter models, these agree with the formulas given in the text.

26. The Jukes-Cantor model can be presented in a different form as a 2×2 Markov model. Let q_t represent the fraction of sites that agree between the ancestral sequence and the descendant sequence at time t, and p_t the fraction that differ, so $q_0 = 1$ and $p_0 = 0$. Assume that the instantaneous rate at which base substitutions occurs is α, and that each of the 3 possible base substitutions is equally likely. Then

$$\begin{pmatrix} q'(t) \\ p'(t) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} \\ \alpha & 1-\frac{\alpha}{3} \end{pmatrix} \begin{pmatrix} q(t) \\ p(t) \end{pmatrix}, \quad \begin{pmatrix} q(0) \\ p(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

    a. Explain why each entry in the matrix has the value it does. (Observe that $1 - \frac{\alpha}{3} = (1-\alpha) + \frac{2\alpha}{3}$.)

    b. Compute the stable distribution of the model by finding the eigenvector with eigenvalue 1.

    c. Find the other eigenvalue and eigenvector for the matrix.

    d. Use (b) and (c), together with the initial conditions to give a formula for $q(t)$ and $p(t)$ as functions of time.

27. Show equation (6.12) implies that $\mathbf{p}$ is a stable base distribution for M.

28. Show that a time reversible rate matrix Q can be expressed by the formula (6.14).

29. Suppose Q is a rate matrix.

    a) Show that if $\mathbf{p}Q = \lambda\mathbf{p}$ and $M(t) = e^{Qt}$ for some t, then $\mathbf{p}M(t) = e^{\lambda t}\mathbf{p}$.

    b) From a) deduce that if $\mathbf{p}Q = 0$, then $\mathbf{p}$ is a stable distribution for all $M(t)$.

30. Suppose Q is a time-reversible rate matrix with stable base distribution $\mathbf{p}$.

    a) Explain why replacing Q with any positive scalar multiple cQ can lead to exactly the same joint distributions on any tree, if edge lengths are adjusted appropriately. Why is this change equivalent to using a new time scale?

b) In light of part (a), it is sometimes convenient to choose a specific normalization of Q. Explain why $-\mathrm{Tr}(\mathrm{diag}(\mathbf{p})Q)$ gives the instantaneous rate of substitutions for the model, and why we can always rescale Q so this is 1. Here $\mathrm{Tr}(M) = \sum_{i=1}^{n} M_{i,i}$ denotes the trace of a matrix M.

31. Show that a time-reversible Markov matrix M with a stable distribution vector $\mathbf{p}$ which has all positive entries must have a full set of real eigenvalues and eigenvectors. (Hint: Show $\mathrm{diag}(\mathbf{p})^{1/2} M \,\mathrm{diag}(\mathbf{p})^{-1/2}$ is symmetric.)

32. The general Markov model is not time reversible, but has a related weaker property.

    a. Show it is not time reversible by giving a specific $\mathbf{p}$, M that do not satisfy equation (6.12).

    b. Show that with mild conditions on $\mathbf{p}$, M, there exist $\tilde{\mathbf{p}}$, $\tilde{M}$ so that

    $$\mathrm{diag}(\mathbf{p})M = \tilde{M}^T \,\mathrm{diag}(\tilde{\mathbf{p}}).$$

    Thus, by changing parameter values for the general Markov model, we can change our viewpoint as to what is ancestral and what is descendant.

    c. Explain the connection between part (b) and Bayes Theorem.