2.3 Markov models of sequence evolution

Now that we have the Markov chain machinery under our belts, let’s return to the question of modeling sequence evolution. The process of substitution at a single site in a nucleotide sequence can be modeled as a Markov chain, where each state represents a single nucleotide. The transition probability, \( P_{jk} \), is the probability that nucleotide \( j \) will be replaced by nucleotide \( k \) in one time step. Similarly, Markov chains can be constructed to model the evolution of amino acid sequences. Although in principle Markov models of sequence evolution are general and can be applied to nucleotide sequences and to amino acid sequences in exactly the same way, in practice working with a twenty-letter alphabet poses challenges that do not arise with a four-letter alphabet. In addition, the biophysical properties of the amino acids are more varied than those of the nucleotides. For these reasons, the Markov chain framework is applied somewhat differently in amino acid sequence models. For the moment, we will focus on nucleotide models and postpone amino acid models until later in the course.

Markov models of sequence substitution are used to answer a wide range of questions that arise in molecular evolution, including correcting for multiple substitutions at the same site, simulating sequence evolution, estimating rates of evolution, deriving substitution scoring matrices, and estimating the likelihood of observing a pair of aligned nucleotides, given a phylogenetic model.

The simplest Markov model of sequence evolution for DNA is the Jukes-Cantor model\(^3\), which assumes that all substitutions (\( A \rightarrow C, A \rightarrow G, A \rightarrow T, C \rightarrow A \ldots \)) are equally probable and occur at a rate, \( \alpha \). Since DNA sequences are made up of four nucleotides, there are three possible substitutions for any given base. Thus, the overall rate of substitution is \( \lambda = 3\alpha \). That is, \( \lambda \) is the probability that a given nucleotide will be replaced by some other nucleotide in one time step. The probability that the nucleotide remains unchanged is \( 1 - 3\alpha \). A graphical representation of this model is shown in Fig. 2.1. The transition probability matrix for this Markov model is:

\[
\begin{bmatrix}
A & G & C & T \\
A & 1-3\alpha & \alpha & \alpha & \alpha \\
G & \alpha & 1-3\alpha & \alpha & \alpha \\
C & \alpha & \alpha & 1-3\alpha & \alpha \\
T & \alpha & \alpha & \alpha & 1-3\alpha
\end{bmatrix}
\]

The rate, \( \alpha \), of each possible substitution is an explicit parameter of the Jukes-Cantor model. In addition, the frequencies of A’s, G’s, C’s and T’s are implicitly specified by the model, since this is determined by the stationary distribution. The stationary distribution

Chapter 2  Sequence Evolution Models

Figure 2.1: Jukes Cantor substitution model

of this Markov chain is \( \varphi^* = (0.25, 0.25, 0.25, 0.25) \). (Verify that this is so by checking that \( \varphi^* = \varphi^*P \)).

Nucleotide substitution models can be made more realistic in two directions. First, the assumption that all substitutions occur at the same rate can be relaxed. Second, the specification of the rates can be adjusted to yield a non-uniform stationary distribution, since the assumption that all four bases have the same frequency (\( \varphi_A = \varphi_C = \varphi_G = \varphi_T \)) is unlikely to hold in most data sets.

2.3.1 Non-uniform transition probabilities

The Kimura 2 Parameter (K2P) model assumes that transitions and transversions occur at different rates. A transition is the substitution of a purine for another purine or a pyrimidine for another pyrimidine. A transversion is the substitution of a purine for a pyrimidine or a pyrimidine for a purine. Recall that the pyrimidines, including cytosine and thymine, are nucleotides with a ring with six elements. The purines, including adenine and guanine, have a pyrimidine ring fused to a five-sided imidazole ring. It make sense that transversions would proceed at a different rate than transitions, since substituting a purine with a pyrimidine, or vice versa, involves a greater change in size and shape than a substitution of two nucleotides from the same class.

The transition matrix for the K2P model is

\[
\begin{bmatrix}
A & G & C & T \\
A & 1-\alpha-2\beta & \alpha & \beta & \beta \\
G & \alpha & 1-\alpha-2\beta & \beta & \beta \\
C & \beta & \beta & 1-\alpha-2\beta & \alpha \\
T & \beta & \beta & \alpha & 1-\alpha-2\beta \\
\end{bmatrix},
\]
where $\alpha$ is the rate of transitions and $\beta$ is the rate of transversions. In this model, the overall substitution rate is $\lambda = \alpha + 2\beta$, since of the three possible substitutions for any given base, one is a transition and two are transversions. Like the Juke Cantor model, the K2P model has a uniform stationary distribution, $\varphi^* = (0.25, 0.25, 0.25, 0.25)$. (Work through the algebra to convince yourself that this is true.)

### 2.3.2 Non-uniform stationary distributions

A stationary distribution with uniform base frequencies is not a realistic model for the many genomes in which the G+C content deviates from 50%. The 1981 Felsenstein (F81) model allows for an arbitrary stationary distribution, $\varphi^* = (\varphi^*_A, \varphi^*_C, \varphi^*_G, \varphi^*_T)$, where $\varphi^*_A \neq \varphi^*_C \neq \varphi^*_G \neq \varphi^*_T$. Like the Jukes-Cantor model, the F81 model has a single scaling parameter, $\alpha$, and does not make a distinction between transitions and transversions. The F81 transition matrix is:

$$
\begin{bmatrix}
A & 1-\alpha (\varphi^*_C + \varphi^*_G + \varphi^*_T) & \alpha \varphi^*_G & \alpha \varphi^*_C & \alpha \varphi^*_T \\
G & \alpha \varphi^*_A & 1-\alpha (\varphi^*_A + \varphi^*_C + \varphi^*_T) & \alpha \varphi^*_C & \alpha \varphi^*_T \\
C & \alpha \varphi^*_A & \alpha \varphi^*_G & 1-\alpha (\varphi^*_A + \varphi^*_G + \varphi^*_T) & \alpha \varphi^*_T \\
T & \alpha \varphi^*_A & \alpha \varphi^*_G & \alpha \varphi^*_C & 1-\alpha (\varphi^*_A + \varphi^*_C + \varphi^*_G)
\end{bmatrix}
$$

### 2.3.3 More general models

The Hasegawa, Kishino, Yano (HKY) model combines both innovations. It allows for different rates for transitions and transversions and an arbitrary stationary distribution, $\varphi^* = (\varphi^*_A, \varphi^*_C, \varphi^*_G, \varphi^*_T)$. The HKY transition matrix is:

$$
\begin{bmatrix}
A & 1-\beta \varphi^*_G & \alpha \varphi^*_G & \beta \varphi^*_C & \beta \varphi^*_T \\
G & \alpha \varphi^*_A & 1-\beta \varphi^*_A & \beta \varphi^*_C & \beta \varphi^*_T \\
C & \beta \varphi^*_A & \beta \varphi^*_G & 1-\beta \varphi^*_A-\beta (\varphi^*_A+\varphi^*_G) & \alpha \varphi^*_T \\
T & \beta \varphi^*_A & \beta \varphi^*_G & \alpha \varphi^*_C & 1-\beta \varphi^*_A-\beta (\varphi^*_A+\varphi^*_G)
\end{bmatrix}
$$

The General Time Reversible (GTR) model is an even more general model that allows a different rate for each of the six possible substitutions and an arbitrary stationary distribution, $\varphi^* = (\varphi^*_A, \varphi^*_C, \varphi^*_G, \varphi^*_T)$ The GTR transition matrix is:

$$
\begin{bmatrix}
A & 1-\beta \varphi^*_G-\gamma \varphi^*_T & \alpha \varphi^*_G & \beta \varphi^*_C & \gamma \varphi^*_T \\
G & \alpha \varphi^*_A & 1-\beta \varphi^*_A-\delta \varphi^*_C & \delta \varphi^*_C & \epsilon \varphi^*_T \\
C & \beta \varphi^*_A & \delta \varphi^*_G & 1-\beta \varphi^*_A-\delta \varphi^*_C-\eta \varphi^*_T & \eta \varphi^*_T \\
T & \gamma \varphi^*_A & \epsilon \varphi^*_G & \eta \varphi^*_C & 1-\gamma \varphi^*_A-\epsilon \varphi^*_C-\eta \varphi^*_G
\end{bmatrix}
$$
All of these models are discussed in greater detail in various molecular evolution textbooks; see, for example, Li’s *Molecular Evolution*, (Sinauer Associates, 1997).

### 2.3.4 Model Selection

In deciding which model to use for a particular data set, we face a trade-off that arises with many statistical models. More general models with more parameters provide a more accurate representation of the underlying evolutionary process. However, with more complex models, more data is required to estimate the parameter values and the danger of overfitting the parameters is greater.

Analyses of alignments of present-day sequences suggest that, in many sequence families, the rate of change varies from site to site. This is typically addressed by assuming that sequence substitution in a given family can be captured by a single model with a small number of rate categories. For example, one might model substitution in a given family using the Jukes Cantor model with four rates, \((\alpha_1, \alpha_2, \alpha_3, \alpha_4)\). For each site, \(i\), maximum likelihood estimation is used to estimate probabilities \((p_1(i), p_2(i), p_3(i), p_4(i))\), where \(p_r(i)\) is the probability that site \(i\) is evolving at rate \(\alpha_r\). Ziheng Yang discusses this approach in his textbook *Computational Molecular Evolution* (Oxford University Press, 2006).

These models do not allow for changes in rate or in GC-content over time. Developing models to account for temporal changes in rate or nucleotide composition is currently an active area of research.

### 2.4 Two applications of DNA substitution models

There are many applications of Markov models of sequence substitution. Here we demonstrate how DNA substitution models can be used to estimate the likelihood of observing a pair of aligned nucleotides, given a phylogenetic model, and to correct for multiple substitutions. In future lectures, we will use an amino acid substitution model to derive a scoring matrix.

#### 2.4.1 The likelihood of a pair of aligned nucleotides

First, let’s consider the problem of estimating the likelihood of a pair of aligned sequences. This problem arises in maximum likelihood approaches to estimating a phylogenetic tree. Maximum likelihood estimation (MLE) is a general method for estimating parameters of a model. It is based on the assumption that the observed data is best explained by the model that maximizes its likelihood; that is, the model for which the probability of the data is greatest. Given a parameterized model, the parameter values are estimated by determining the values that maximize the probability of the data.
2.4 Two applications of DNA substitution models

Figure 2.2: A hypothetical evolutionary scenario: (a) A pairwise alignment of two sequences with residues \( x = G \) and \( y = C \) at site \( i \). (b) These residues in present-day nucleotide sequences have been diverging from a common, unknown ancestral nucleotide, \( z \), for a period of time, \( t \). (c) In the Jukes Cantor model, the calculation of the probability of seeing a given nucleotide at site \( i \) after elapsed time, \( t \), can be reduced to two cases: either the nucleotides in the ancestral and present-day sequences are the same (left, Equation 2.29) or they are not (right, Equation 2.30).

In the context of phylogeny estimation, the observed data is a set of \( k \) aligned sequences. The model has two components: a Markov model of sequence substitution and a rooted, binary tree with \( k \) leaves. The likelihood is the probability of observing the multiple alignment, under the assumption that the sequences evolved along the branches of the tree, sustaining mutations according to the rates specified by the substitution model. For a fixed tree topology, the branch lengths and the substitution rates are estimated by maximizing the probability of the multiple sequence alignment.

We demonstrate this calculation for the case where \( k = 2 \). Suppose we have two residues, \( x \) and \( y \), that are the descendants of an ancestral nucleotide, \( z \), and that time \( t \) has elapsed since their divergence. Fig. 2.2 illustrates this situation for the case where \( x \) is a guanine and \( y \) is a cytosine. The probability of observing \( x \) aligned with \( y \) is the product of \( p_{zx}(t) \), the probability of observing a \( z \) at time \( t \), and \( p_{zy}(t) \), the probability of observing a \( y \) at time \( t \), given that the ancestral residue was \( z \). Since the base in the ancestral sequence is unknown, we estimate the probability by the weighted sum over all possible values of \( z \):

\[
\Pr \left( x \mid y, t, \alpha \right) = \sum_{z \in \{A, C, G, T\}} p_z \ p_{zx}(t) \ p_{zy}(t),
\]

(2.25)

where \( p_z \) is an estimation of the frequency of \( z \) in the ancestral sequence. For example, the likelihood of observing a guanine in one sequence and a cytosine in the other is

\[
\Pr \left( G \mid t, \alpha \right) = p_A \ p_{AG}(t) + p_C \ p_{CG}(t) + p_G \ p_{GG}(t) + p_T \ p_{TG}(t).
\]

(2.26)
In order to estimate the probability of observing $x$ aligned with $y$, given that a time interval $t$ has elapsed since they diverged from their common ancestor, we need a way to estimate $p_{xx}(t)$ and $p_{zy}(t)$. Here, we derive expressions for the probability $p_{xx}(t)$ under the assumption that the sequence is evolving according to the Jukes Cantor model. The Jukes Cantor transition probability matrix,

$$
\begin{bmatrix}
1 & -3\alpha & \alpha & \alpha & \alpha \\
\alpha & 1-3\alpha & \alpha & \alpha & \alpha \\
\alpha & \alpha & 1-3\alpha & \alpha & \alpha \\
\alpha & \alpha & \alpha & 1-3\alpha & \alpha \\
\end{bmatrix}
$$

is defined in terms of an instantaneous substitution rate, $\alpha$. If the duration of a single time step in this Markov chain is $\delta t$, $\delta t \ll \frac{1}{\alpha}$, then the probability of a substitution between a given pair of nucleic acids in a single time step is $\alpha \delta t$. We can define a transition probability matrix for the Jukes Cantor Markov model with this time interval as follows:

$$
\begin{bmatrix}
1-3\alpha \delta t & \alpha \delta t & \alpha \delta t & \alpha \delta t \\
\alpha \delta t & 1-3\alpha \delta t & \alpha \delta t & \alpha \delta t \\
\alpha \delta t & \alpha \delta t & 1-3\alpha \delta t & \alpha \delta t \\
\alpha \delta t & \alpha \delta t & \alpha \delta t & 1-3\alpha \delta t \\
\end{bmatrix}
$$

We use this transition matrix to derive an expression describing how changes accumulate at site $i$ over a period of time $t$. First, we consider the event of observing, for example, an A at site $i$ after a single time step; i.e., at time $t + \delta t$. This event can occur in two ways: either site $i$ contained an A at time $t$ and no substitution occurred during the most recent time step or site $i$ contained some other nucleotide at time $t$ and a substitution resulted in an A one time step later. Accounting for both of these scenarios, the probability of observing A at time $t + \delta t$ is

$$
\varphi_A(t + \delta t) = (1-3\alpha \delta t) \varphi_A(t) + \alpha \delta t \varphi_C(t) + \alpha \delta t \varphi_G(t) + \alpha \delta t \varphi_T(t)
$$

where $\varphi_x(t)$ is the probability of observing nucleotide $x$ (i.e. of being in state $E_x$) at time $t$. Since in the stationary distribution of the Jukes Cantor model, all nucleotides have the same frequency, we can combine the 2nd, 3rd and 4th terms, yielding

$$
\varphi_A(t + \delta t) = (1-3\alpha \delta t) \varphi_A(t) + \alpha \delta t [1-\varphi_A(t)].
$$

Here, the first term gives the probability that the residue at site $i$ at time $t$ was an A and no substitution occurred. The second term is the probability that the residue at time $t$ was not an A and a substitution did occur, replacing that residue with A. Since the model is symmetric, this equation applies equally well to $C, G$ or $T$. We can therefore rewrite the equation using the parameter $x$, where $x \in \{A, C, G, T\}$, and combine terms to obtain

$$
\varphi_x(t + \delta t) = \varphi_x(t) + (1-4 \varphi_x(t)) \alpha \delta t.
$$  \hspace{1cm} (2.27)
Having obtained an expression for the probability of observing a given nucleotide \((x)\) after one time step, we next need to derive an expression for the probability of observing \(x\) after a longer time interval. Subtracting \(\varphi_x(t)\) from both sides of Equation 2.27 and some algebraic manipulation yields

\[
\frac{\varphi_x(t + \delta t) - \varphi_x(t)}{\delta t} = (1 - 4 \varphi_x(t)) \alpha.
\]

Taking the limit as \(\delta t \to 0\), we obtain a differential equation

\[
\frac{d\varphi_x(t)}{dt} = (1 - 4 \varphi_x(t)) \alpha
\]

(2.28)

that we can use to obtain an expression for the probability of observing nucleotide \(x\) at site \(i\) after an arbitrary time interval \(t\). This differential equation has a standard form \((f'(t) = a - bf(t))\) with a known solution, from which we obtain

\[
\varphi_x(t) = \frac{1}{4} + \left(\varphi_x(0) - \frac{1}{4}\right) e^{-4\alpha t}.
\]

**NOTE:** The standard form linear differential equation \(\frac{df(x)}{dx} = q(x) - p(x)f(x)\) can be solved with the formula \(f(x) = g(x)^{-1} \int (g(x)q(x)dx) + C \cdot g(x)^{-1}\), where \(g(x) = e\int p(x)dx\).

In our case, the differential equation is \(\frac{d\varphi_x(t)}{dt} = \alpha - 4\alpha \varphi_x(t)\) so when substituting in the standard form: \(x = t\), \(f(x) = \varphi_x(t)\), \(q(t) = \alpha\), and \(p(t) = 4\alpha\). Solving, \(g(t) = e\int p(t)dt\) yields \(e^\int 4\alpha dt = e^{4\alpha t}\). From the standard form solution, we get

\[
\varphi_x(t) = \frac{1}{4} + \left(\varphi_x(0) - \frac{1}{4}\right) e^{-4\alpha t}.
\]

We now have an expression that gives the probability of observing nucleotide \(x\) in terms of initial state probability, \(\varphi_x(0)\). So how to solve the probability of observing nucleotide \(x\), given the ancestral nucleotide \(z\)? We have two cases. Either the nucleotide at this site in the ancestral sequence was also \(x\) (i.e., \(\varphi_x(0) = 1\)) or the nucleotide at this site in the ancestral sequence was some residue other than \(x\) (i.e., \(\varphi_x(0) = 0\)). Substituting 1 for \(\varphi_x(0)\) in the
expression for $\varphi_x(t)$, we can derive an expression for the probability that the present-day residue is the same as the ancestral nucleotide after time $t$:

$$p_{xx}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}. \tag{2.29}$$

Similarly, setting $\varphi_x(0) = 0$ in the expression for $\varphi_x(t)$ gives the probability that the present-day nucleotide differs from the ancestral residue after time $t$:

$$p_{zx}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}. \tag{2.30}$$

Using Equations 2.29 and 2.30, we can now derive an expression, in terms of $\alpha$ and $t$, for the likelihood of observing $G$ aligned with $C$. Substituting the right hand sides of Equations 2.29 and 2.30 into Equation 2.26, we obtain the likelihood for observing $G$ aligned with $C$

$$\Pr(G | C | t, \alpha) = \frac{1}{2} \left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right)^2 + \frac{1}{2} \left( \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right) \cdot \left( \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \right),$$

assuming that $p_z = \frac{1}{4}$ for all $z$.

Equations 2.29 and 2.30 can be used to derive the more general $\Pr(x | y | t, \alpha)$ and $\Pr(x | x | t, \alpha)$, the probabilities of aligning two nucleotides when the nucleotides mismatch and match, respectively:

$$\Pr(x | y | t, \alpha) = \frac{1}{4} \sum_{z \in \{A,C,G,T\}} p_{zx}(t)p_{zy}(t) \Pr(x | t, \alpha) = \frac{1}{4} \sum_{z \in \{A,C,G,T\}} p_{zx}(t)^2. \tag{2.31}$$

When the aligned sequences are a mismatch, there are two cases: (1) two ancestral nucleotides will the same as one residue ($x$ and then $y$) and differ from the other and (2) two ancestral nucleotides will differ from both aligned residues. Thus, the probability of aligning two different nucleotides is:

$$\Pr(x | y | t, \alpha) = \frac{1}{4} \left( 2p_{xx}(t)p_{zx}(t) + 2p_{zx}(t)^2 \right)$$

$$= \frac{1}{2} p_{zx}(t) \left( p_{xx}(t) + p_{zx}(t) \right)$$

$$= \frac{1}{8} \left( 1 - e^{-4\alpha t} \right) \left( \frac{1}{4} (1 + 3e^{-4\alpha t}) + \frac{1}{4} (1 - e^{-4\alpha t}) \right)$$

$$= \frac{1}{32} \left( 1 - e^{-4\alpha t} \right) (2 + 2e^{-4\alpha t})$$

$$= \frac{1}{16} \left( 1 - e^{-4\alpha t} \right) (1 + e^{-4\alpha t})$$

$$= \frac{1}{16} \left( 1 - e^{-8\alpha t} \right).$$
2.4 Two applications of DNA substitution models

When the aligned sequences match, there are two cases: (1) one ancestral nucleotide will be the same as both residues and (2) three ancestral nucleotides will differ from both aligned residues. Thus, the probability of aligning two of the same nucleotides is:

\[
\Pr\left( x \mid x, t, \alpha \right) = \frac{1}{4} \left( p_{xx}(t)^2 + 3p_{xx}(t)^2 \right) \\
= \frac{1}{4} \left( \frac{1}{16} (1 + 3e^{-4\alpha t})^2 + \frac{3}{16} (1 - e^{-4\alpha t})^2 \right) \\
= \frac{1}{64} \left( 1 + 6e^{-4\alpha t} + 9e^{-8\alpha t} + 3 - 6e^{-4\alpha t} + 3e^{-8\alpha t} \right) \\
= \frac{1}{64} \left( 4 + 12e^{-8\alpha t} \right) \\
= \frac{1}{16} \left( 1 + 3e^{-8\alpha t} \right).
\]

We now have an expression for the probability of observing nucleotide \( x \) aligned with nucleotide \( y \) that depends on two parameters: the branch length, \( t \), and the substitution rate, \( \alpha \). These parameter values are then estimated by finding the values of \( t \) and \( \alpha \) that maximize \( \Pr\left( x \mid y, t, \alpha \right) \).

This approach can be expanded to values of \( k \) greater than two by nesting multiple expressions with the same form as the right hand side of Equation 2.25. Under the assumption of positional independence, the likelihood for multiple sites is simply the product of the likelihoods for each site, individually.

2.4.2 Correcting for multiple substitutions.

Another problem that arises in molecular evolution is estimating the amount of sequence divergence between a pair of sequences, \( s_1 \) and \( s_2 \). A simple approach to estimating the distance between \( s_1 \) and \( s_2 \) would be to count the number of positions that are not identical in the pairwise alignment of \( s_1 \) and \( s_2 \). If only a few changes have occurred, then the observed number of mismatches may, in fact, be the actual number of substitutions. However, as the divergence increases, so does the probability of two or more substitutions at the same site. In this case, the number of observed changes will underestimate the actual divergence, as shown in Fig. 2.3.

Recall that the Jukes-Cantor model assumes that all substitutions (\( A \rightarrow C, A \rightarrow G, A \rightarrow T, C \rightarrow A \ldots \)) are equally likely and occur at a rate \( \alpha \), resulting in an overall rate of substitution \( \lambda = 3\alpha \). Suppose that we have an ungapped\(^4\) pairwise alignment of length \( n \) of two nucleotide sequences, \( s_1 \) and \( s_2 \), that disagree at \( m \) positions. We wish to estimate the number of substitutions that actually occurred over \( t \), the time interval that elapsed since they diverged from a common ancestor.

\(^4\)None of the substitution models we have discussed account for insertions and deletions.
Here, we use the Jukes-Cantor model to derive a more accurate estimate of the number of substitutions. If we assume a constant rate of substitution, $\lambda$, in both lineages then the expected number of substitutions per site is

$$P_s = 2\lambda t = 6\alpha t.$$ 

Since both $\alpha$ and $t$ are unknown, we estimate the expected number of substitutions from the frequency of mismatches in the current alignment. Given a Markov model of sequence substitution, we can use the observed frequency of mismatches to estimate $2\lambda t$ using the following strategy:

First, using the expressions for $p_{xx}(t)$ and $p_{zx}(t)$ that we derived in the previous section (Eqns. 2.29 and 2.30), we estimate the frequency of mismatches as a function of $\alpha t$,

$$P_m = f(\alpha t).$$

We do this by estimating $P_M$, the frequency of matches, and subtracting to obtain $P_m = 1 - P_M$. Next, we invert this function to obtain an expression for the expected number of substitutions per single site in terms of the number of mismatches per site:

$$\alpha t = f^{-1}(P_m).$$

The underlying frequency of mismatches can be approximated by the observed frequency of mismatches, $\frac{m}{n}$, yielding an equation of the form

$$\alpha t \approx f^{-1}\left(\frac{m}{n}\right).$$
From this, we obtain an approximation for the frequency of substitutions:

\[
P_s = 6\alpha t \\
\approx 6f^{-1}(\frac{m}{n}).
\]

Now we apply this strategy to obtain an estimate of the number of substitutions that occurred assuming that sequences are evolving according to the Jukes Cantor model. First, we derive an expression for the probability of observing a match; for example, for observing two adenines aligned at site \(i\). Given two sequences evolving independently from a common ancestral sequence, the probability that both sequences will have an \(x\) at site \(i\) is

\[
P_M = [p_{Ax}(t)]^2 + [p_{Tx}(t)]^2 + [p_{Cx}(t)]^2 + [p_{Gx}(t)]^2,
\]

where \(t\) is the elapsed time since their divergence. One of these terms will represent the case where the ancestral residue is \(x\); the three others will represent the cases where the ancestral residue is not \(x\). Replacing the one term with Equation 2.29 and the remaining three terms with Equation 2.30, this reduces to

\[
P_M = \left[\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right]^2 + 3\left[\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right]^2.
\]

The first term gives the probability of observing \(x\)'s in both sequences if the ancestral nucleotide was also \(x\). The second term represents the cases where the ancestral nucleotide was not an \(x\). By expanding the squared quantities and combining terms, we obtain

\[
P_M = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}. \quad (2.32)
\]

The probability of observing a mismatch at site \(i\) is simply \(1 - P_M\) or

\[
P_m = \frac{3}{4} \left(1 - e^{-8\alpha t}\right). \quad (2.33)
\]

We solve the above equation to obtain an expression for \(\alpha t\) in terms of \(P_m\):

\[
\alpha t = -\frac{1}{8} \ln \left(1 - \frac{4}{3}P_m\right).
\]

Multiplying both sides of the equation by 6 yields \(6\alpha t\), which is the expected frequency of substitutions per site in terms of the probability of observing a mismatch:

\[
P_s = -\frac{3}{4} \ln \left(1 - \frac{4}{3}P_m\right).
\]
$P_m$ can be estimated by the observed frequency of mismatches, allowing us to obtain an estimate in terms of the fraction of sites with an observable difference:

$$P_s \approx -\frac{3}{4} \ln \left(1 - \frac{4}{3} \frac{m}{n}\right).$$

Multiplying by $n$ yields an estimate of the expected number of substitutions that actually occurred in the entire alignment of length $n$:

$$-\frac{3}{4} \ln \left(1 - \frac{4}{3} \frac{m}{n}\right) \cdot n. \quad (2.34)$$

For example, if we observe mismatches at 100 sites in a pairwise nucleotide sequence alignment of length 1,000, then the Jukes-Cantor model predicts that the actual number of substitutions per site is 0.107 or 107 substitutions.

### 2.4.3 Applications with the K2P model

In previous sections, we used the Jukes Cantor model to derive expressions for the probability of observing a given nucleotide at site $i$ in a present day sequence (Equations 2.29 and 2.30). We then used these equations to obtain an estimate for the number of substitutions that occurred, given the number of mismatches observed in an ungapped alignment (Equation 2.34). Analogous equations can be derived for the K2P model, which assumes that transitions and transversions occur at different rates.

In this case, we will use the observation for continuous Markov chains that the continuous rate probabilities can be computed from the derivative transition matrix $P'(t)$, which can be calculated from the product of the transition matrix $P(t)$ and the instantaneous rates at $t = 0$. Specifically, $P'(t) = P(t)P'(0)$

For the K2P model, the transition matrix $P(t)$ is:

$$
\begin{bmatrix}
A & G & C & T \\
A & p_{AA}(t) & p_{AG}(t) & p_{AC}(t) & p_{AT}(t) \\
G & p_{GA}(t) & p_{GG}(t) & p_{GC}(t) & p_{GT}(t) \\
C & p_{CA}(t) & p_{CG}(t) & p_{CC}(t) & p_{CT}(t) \\
T & p_{TA}(t) & p_{TG}(t) & p_{TC}(t) & p_{TT}(t)
\end{bmatrix}
$$

The probability of observing different nucleotides at site $i$ in the ancestral and present-day sequences depends on whether the nucleotides belong to the same class or to different classes. If the ancestral nucleotide $z$ belongs to the same class as the present-day nucleotide (i.e., there was a transition), I denote the present-day nucleotide as $x$ and the probability of $z \to x$ over time $t$ as $p_{zx}(t)$. If the ancestral nucleotide $z$ belongs to a different class as the present-day nucleotide (i.e., there was a transversion), I denote the present-day nucleotide as $y$ and the probability of $z \to y$ over time $t$ as $p_{zy}(t)$. Under the K2P model, the probability
that the ancestral nucleotide \( z \) at site \( i \) is the same as the present-day nucleotide after time \( t \) is \( p_{zz}(t) = 1 - p_{zx}^s(t) - 2p_{zy}^v(t) \). The transition matrix \( P(t) \) can then be written as:

\[
\begin{bmatrix}
A & 1 - p_{zx}^s(t) - 2p_{zy}^v(t) \\
A & p_{zx}^s(t) \\
G & 1 - p_{zx}^s(t) - 2p_{zy}^v(t) \\
G & p_{zx}^s(t) \\
C & 1 - p_{zx}^s(t) - 2p_{zy}^v(t) \\
C & p_{zx}^s(t) \\
T & 1 - p_{zx}^s(t) - 2p_{zy}^v(t) \\
T & p_{zx}^s(t)
\end{bmatrix}
\]

(2.35)

The derivative matrix \( P'(t) \) is found by taking the derivative of each entry with respect to \( t \), yielding:

\[
\begin{bmatrix}
A & -p_{zx}^s(t) - 2p_{zy}^v(t) \\
A & p_{zx}^s(t) \\
G & -p_{zx}^s(t) - 2p_{zy}^v(t) \\
G & p_{zx}^s(t) \\
C & -p_{zx}^s(t) - 2p_{zy}^v(t) \\
C & p_{zx}^s(t) \\
T & -p_{zx}^s(t) - 2p_{zy}^v(t) \\
T & p_{zx}^s(t)
\end{bmatrix}
\]

Setting \( p_{zx}^s(t) = \alpha \) and \( p_{zy}^v(t) = \beta \), then the derivative transition matrix at \( t = 0 \), \( P'(0) \) is

\[
\begin{bmatrix}
A & \alpha - \alpha - 2\beta & \alpha & \beta \\
A & \alpha & \alpha - 2\beta & \beta \\
G & \beta & \beta & \alpha - 2\beta \\
G & \beta & \beta & \alpha \\
C & \beta & \beta & \alpha - 2\beta \\
C & \beta & \beta & \alpha \\
T & \beta & \beta & \alpha - 2\beta \\
T & \beta & \beta & \alpha
\end{bmatrix}
\]

(2.36)

Our goal is to solve the derivatives \( \frac{dp_{zx}^s(t)}{dt} \) and \( \frac{dp_{zy}^v(t)}{dt} \). We do this by taking any transition and transversion entry, respectively, in \( P'(t) = P(t) \times P'(0) \) (i.e., \( p'_{AC}(t) \) and \( p'_{AC}(t) \)) and solving using the method for standard form linear differential equations, as we did in the example with Jukes-Cantor. Looking at matrices 2.35 and 2.36, we see that these derivatives are

\[
\frac{dp_{zx}^s(t)}{dt} = \alpha \left(1 - p_{zx}^s(t) - 2p_{zy}^v(t)\right) - (\alpha + 2\beta)p_{zx}^s(t) + \beta p_{zy}^v(t) + \beta p_{zy}^v(t)
\]

\[
= \alpha - \alpha p_{zx}^s(t) - 2\alpha p_{zy}^v(t) - \alpha p_{zx}^s(t) - 2\beta p_{zx}^s(t) + 2\beta p_{zy}^v(t) + \beta p_{zy}^v(t)
\]

\[
= \alpha - 2(\alpha + \beta)p_{zx}^s(t) - 2(\alpha - \beta)p_{zy}^v(t)
\]

\[
\frac{dp_{zy}^v(t)}{dt} = \beta \left(1 - p_{zx}^s(t) - 2p_{zy}^v(t)\right) + \beta p_{zx}^s(t) - (\alpha + 2\beta)p_{zy}^v(t) + \alpha p_{zy}^v(t)
\]

\[
= \beta - \beta p_{zx}^s(t) - 2\beta p_{zy}^v(t) + \beta p_{zx}^s(t) - \alpha p_{zy}^v(t) - 2\beta p_{zy}^v(t) + \alpha p_{zy}^v(t)
\]

\[
= \beta - 4\beta p_{zy}^v(t).
\]

Since \( \frac{dp_{zy}^v(t)}{dt} \) only depends on \( p_{zy}^v(t) \), we will solve the transversion case first. In fact, this is exactly the same form as differential equation 2.28, which we solved for the JC
model. Thus, \( p_{zy}^v(t) = \frac{1}{4} + C \cdot e^{-4\beta t} \) and \( C = -\frac{1}{4} \) since \( p_{zy}^v(0) = 0 = \frac{1}{4} + C \). Therefore, the probability of observing a transversion substitution at site \( i \) in a descendant sequence after elapsed time \( t \) is

\[
p_{zy}^v(t) = \frac{1}{4} - \frac{1}{4} e^{-4\beta t}. \tag{2.37}
\]

We can then substitute this value in \( \frac{dp_{zy}^v(t)}{dt} \) to get

\[
\frac{dp_{zy}^v(t)}{dt} = \frac{\alpha - 2 (\alpha + \beta) e^{2(\alpha + \beta)t} - \frac{1}{2} (\alpha - \beta) (1 - e^{-4\beta t})}{\alpha - 2 (\alpha + \beta) e^{2(\alpha + \beta)t} - \frac{1}{2} \alpha + \frac{1}{2} \beta + \frac{1}{2} (\alpha - \beta) e^{-4\beta t}} = \frac{1}{2} (\alpha + \beta) + \frac{1}{2} (\alpha - \beta) e^{-4\beta t} - 2 (\alpha + \beta) p_{zy}^v(t)
\]

Recall that \( \frac{df(t)}{dt} = g(t) - h(t)f(t) \) is the standard form linear differential equation, which can be solved using the formula \( f(t) = g(t)^{-1} \int g(t)q(t)dt + C \cdot g(t)^{-1} \), where \( g(t) = e^{\int h(t)dt} \). In our case, \( f(t) = p_{zy}^v(t) \), \( q(t) = \frac{1}{4} (\alpha + \beta) + \frac{1}{2} (\alpha - \beta) e^{-4\beta t} \), and \( h(t) = 2(\alpha + \beta) \). Solving \( g(t) = e^{\int h(t)dt} \) yields

\[
g(t) = e^{\int 2(\alpha + \beta)dt} = e^{2(\alpha + \beta)t}.
\]

We then need to solve \( \int g(t)q(t)dt \):

\[
\int g(t)q(t)dt = \int e^{2(\alpha + \beta)t} \left( \frac{1}{2} (\alpha + \beta) + \frac{1}{2} (\alpha - \beta) e^{-4\beta t} \right) dt
\]

\[
= \int \frac{1}{2} (\alpha + \beta) e^{2(\alpha + \beta)t} + \int \frac{1}{2} (\alpha - \beta) e^{-4\beta t} e^{2(\alpha + \beta)t} dt
\]

\[
= \frac{\alpha + \beta}{2} \int e^{2(\alpha + \beta)t} dt + \frac{\alpha - \beta}{2} \int e^{2(\alpha - \beta)t} dt
\]

\[
= \frac{\alpha + \beta}{4(\alpha + \beta)} e^{2(\alpha + \beta)t} + \frac{\alpha - \beta}{4(\alpha - \beta)} e^{2(\alpha - \beta)t}
\]

\[
= \frac{1}{4} \left( e^{2(\alpha + \beta)t} + e^{2(\alpha - \beta)t} \right).
\]

Plugging these values into the standard form solution yields

\[
f(t) = g(t)^{-1} \int g(t)q(t)dt + C \cdot g(t)^{-1}
\]

\[
p_{zy}^v(t) = e^{-2(\alpha + \beta)t} \left( \frac{1}{4} \left( e^{2(\alpha + \beta)t} + e^{2(\alpha - \beta)t} \right) + C \cdot e^{-2(\alpha + \beta)t} \right)
\]

\[
= \frac{1}{4} \left( e^{-2(\alpha + \beta)t} e^{2(\alpha + \beta)t} + e^{-2(\alpha + \beta)t} e^{2(\alpha - \beta)t} \right) + C \cdot e^{-2(\alpha + \beta)t}
\]

\[
= \frac{1}{4} \left( 1 + e^{-4(\alpha + \beta)t} \right) + C \cdot e^{-2(\alpha + \beta)t}.
\]
Solving for $C$, $p_{zx}^s(0) = 0 = \frac{2}{3} + C$ and $C = -\frac{1}{2}$. Therefore, the probability of observing a transition substitution at site $i$ in a descendant sequence after elapsed time $t$ is

$$p_{zx}^s(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t}. \quad (2.38)$$

Finally, Equations 2.37 and 2.38 can be combined to solve $p_{zz}(t) = 1 - p_{zx}^s(t) - 2p_{zy}^v(t)$. The probability of observing the same nucleotide at site $i$ in an ancestral sequence and in a descendant sequence after elapsed time $t$ is

$$p_{zz}(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}. \quad (2.39)$$

This equation is analogous to Equation 2.29.

The likelihood of a pair of aligned nucleotides

When inferring the likelihood of observing a pair of nucleotides aligned at site $i$, we are actually considering a tree with two leaves and inferring the probability that some ancestral nucleotide $z$ at site $i$ evolved independently over time $t$ into the nucleotides observed in the present-day sequences at site $i$. Let $j$ and $k$ be the observed, present-day nucleotides.

The probability of observing $j$ in the one present-day sequence is $p_{zj}(t)$ and the probability of observing $k$ in the other present-day species is $p_{zk}(t)$. The probability of observing $j$ aligned with $k$ is a product of the probabilities $p_{zj}(t)$ and $p_{zk}(t)$. Since the ancestral nucleotide $z$ is unknown, we consider all four possibilities, weighted by the probability of observing $z$:

$$\Pr(j \mid k) = \sum_{z \in \{A, C, G, T\}} p_z p_{zj}(t) p_{zk}(t). \quad (2.40)$$

Under the K2P model, transitions and transversions are weighted differently. So, we’ll consider three cases for the aligned nucleotides $j$ and $k$: (1) they are the same, (2) they are from the same class (i.e., A aligned with G or C aligned with T), or (3) they are from different classes (i.e., G aligned with C).

The aligned nucleotides are the same.

In this case, we can rewrite Eqn. 2.40 as

$$\Pr(j \mid j) = \frac{1}{4} \sum_{z \in \{A, C, G, T\}} p_{zj}(t)^2.$$
Chapter 2  Sequence Evolution Models

Plugging in Eqns. 2.37, 2.38 and 2.39, yields

\[
\Pr\left(\frac{\mathbf{j}}{\mathbf{j}}\right) = \frac{1}{4} \left(\frac{1}{16}(1 + e^{-4\beta t} + 2e^{-2(\alpha+\beta)t})^2 + \frac{1}{16}(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t})^2 + \frac{1}{16}(1 - e^{-4\beta t})\right)
\]

\[
= \frac{1}{4} \left(\frac{1}{16} + \frac{1}{16} + \frac{1}{16}\right)
\]

\[
= \frac{1}{16} (1 + e^{-8\beta t} + 2e^{-4(\alpha+\beta)t})
\]

The aligned nucleotides are in the same class.

Let’s consider aligning A and G, which give us

\[
\Pr\left(\frac{A}{G}\right) = \frac{1}{4}(p_{AA}(t)p_{AG}(t) + p_{GA}(t)p_{GG}(t) + p_{CA}(t)p_{CG}(t) + p_{TA}(t)p_{TG}(t))
\]

Of the four products, two of them are of the form \(p_{zz}(t)p_{zx}(t)\) and two are of the form \(p_{xy}(t)^2\). This is because in two cases, the ancestral nucleotide will be the same as one of the nucleotides in the alignment; since \(j\) and \(k\) are in the same class, the other present-day nucleotide will be a transition. In the other two cases, the ancestral nucleotide will be in a different class from \(j\) and \(k\), and thus represent a transversion. Therefore, the probability of aligning \(j\) and \(k\) when \(j\) and \(k\) are in the same class is

\[
\Pr\left(\frac{\mathbf{j}}{\mathbf{k}}\right) = \frac{1}{4} \left(2p_{zz}(t)p_{zx}(t) + 2p_{xy}(t)^2\right)
\]

Plugging in Eqns. 2.37, 2.38 and 2.39, yields

\[
\Pr\left(\frac{\mathbf{j}}{\mathbf{k}}\right) = \frac{1}{2} \left(p_{zz}(t)p_{zx}(t) + p_{xy}(t)^2\right)
\]

\[
= \frac{1}{32} \left((1 + e^{-4\beta t} + 2e^{-2(\alpha+\beta)t})(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}) + (1 - e^{-4\beta t})^2\right)
\]

\[
= \frac{1}{16} (1 + e^{-8\beta t} + 2e^{-4(\alpha+\beta)t})
\]

The aligned nucleotides are in different classes.

Let’s consider aligning A and C, which give us

\[
\Pr\left(\frac{A}{C}\right) = \frac{1}{4} (p_{AA}(t)p_{AC}(t) + p_{GA}(t)p_{GC}(t) + p_{CA}(t)p_{CC}(t) + p_{TA}(t)p_{TC}(t))
\]

Of the four products, two of them are of the form \(p_{zz}(t)p_{zy}(t)\) and two are of the form \(p_{zx}(t)p_{zy}(t)\). This is because in two cases, the ancestral nucleotide will be the same as one of the nucleotides in the alignment; since \(j\) and \(k\) are in different classes, the other present-day
nucleotide will be a transversion. In the other two cases, the ancestral nucleotide will be in
the same class as one present-day nucleotide (i.e., a transition) but in a different class for
the other present-day nucleotide (i.e., a transversion). Therefore, the probability of aligning
\( j \) and \( k \) when \( j \) and \( k \) are in different classes is

\[
\Pr \left( \frac{j}{k} \right) = \frac{1}{4} \left( 2p_{zz}(t)p_{zy}^{*}(t) + 2p_{zx}(t)p_{zy}(t) \right)
\]

Plugging in Eqns 2.37, 2.38 and 2.39, yields

\[
\Pr \left( \frac{j}{k} \right) = \frac{1}{32} \left( 1 - e^{-4\beta t} \right) \left( 1 + 2e^{-4(\alpha+\beta)t} + e^{-8\beta t} - 2e^{-2(\alpha+\beta)t} \right)
\]

(2.47)

\[
= \frac{1}{16} \left( 1 - e^{-4\beta t} \right) (1 + e^{-4\beta t})
\]

(2.48)

\[
= \frac{1}{16} \left( 1 - e^{-8\beta t} \right)
\]

(2.49)

Correcting for multiple substitutions.

With the JC model, we calculated the expected number of substitutions per site based on
the probability of a mismatch. In the K2P model, we also have two mismatch probabilities,
one for transitions and one for transversions. Recall that Eqns 2.41, 2.44, and 2.47 give the
probabilities of observing a specific alignment, based on whether the nucleotides aligned at
site \( i \) are the the same, in the same class, or in different classes, respectively. We can use
these equations to calculate the probabilities of a match and the two types of mismatches in
general. In particular, there are four ways to have a match, four ways to have a mismatch
within the same class, and 8 ways to have a mismatch across different classes.

Given an alignment of two sequences that have been diverging for time \( t \) from a common
ancestor and have since been evolving according to the K2P model, the probability of
observing a match (i.e., AA, GG, CC, TT) at a given site \( i \) is

\[
P_{M} = \frac{1}{4} \left( 1 + 2e^{-4(\alpha+\beta)t} + e^{-8\beta t} \right).
\]

(2.52)

The probability of observing a transition (i.e., AG, GA, CT, TC) at a given site \( i \) is

\[
P_{m} = \frac{1}{4} \left( 1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t} \right).
\]

(2.53)

The probability of observing a transversion at site \( i \) is given by

\[
P_{v} = \frac{1}{2} \left( 1 - e^{-8\beta t} \right).
\]

(2.54)
Chapter 2  Sequence Evolution Models

The expected number of substitutions is $E = 2\lambda t = 2(\alpha + 2\beta)t$, with the expected number of transitions $E^s = 2\alpha t$ and the expected number of transversions $E^v = 4\beta t$. We can solve for $\beta t$ and $\alpha t$ from Eqn.s 2.53 and 2.54.

$$2P^v_m = 1 - e^{-8\beta t}$$
$$1 - 2P^v_m = e^{-8\beta t}$$
$$-8\beta t = \ln[1 - 2P^v_m]$$
$$\beta t = -\frac{1}{8} \ln[1 - 2P^v_m]$$

$$4P^s_m = 1 - 2e^{-4(\alpha + \beta)t} + 1 - 2P^v_m$$
$$4P^s_m + 2P^v_m - 2 = -2e^{-4(\alpha + \beta)t}$$
$$1 - 2P^s_m - P^v_m = e^{-4(\alpha + \beta)t}$$
$$-4(\alpha + \beta)t = \ln[1 - 2P^s_m - P^v_m]$$
$$\alpha t + \beta t = -\frac{1}{4} \ln[1 - 2P^s_m - P^v_m]$$

$$\alpha t = \frac{1}{8} \ln[1 - 2P^v_m] - \frac{1}{4} \ln[1 - 2P^s_m - P^v_m]$$
$$\alpha t = \frac{1}{8} (\ln[1 - 2P^v_m] - 2 \ln[1 - 2P^s_m - P^v_m]).$$

Given an alignment of length $n$, with $m_s$ transitions and $m_v$ transversions, the expected numbers of transitions and transversions that actually occurred is approximately

$$\frac{1}{4} \left[ \ln \left( 1 - \frac{2m_v}{n} \right) - 2 \ln \left( 1 - \frac{2m_s}{n} - \frac{m_v}{n} \right) \right] \cdot n \quad (2.55)$$

and

$$-\frac{1}{2} \ln \left( 1 - \frac{2m_v}{n} \right) \cdot n, \quad (2.56)$$

respectively. Summing these two quantities, we obtain the expected number of substitutions of all types:

$$-\frac{1}{4} \left[ \ln \left( 1 - \frac{2m_v}{n} \right) + 2 \ln \left( 1 - \frac{2m_s}{n} - \frac{m_v}{n} \right) \right] \cdot n. \quad (2.57)$$