Lecture Notes: Markov models of sequence evolution

Dannie Durand

These notes cover the lectures on Markov chains, Markov models of sequence evolution, and applications of sequence evolution models to problems in molecular evolution (September 22nd through October 1st).

Introduction to Markov chains

At the beginning of the semester, we introduced two simple scoring functions for pairwise alignments:

- a similarity function, that assigns a score of M to matches ($M > 0$), m to mismatches ($m < 0$), and g to indels ($g < 0$) and
- an edit distance, which does not reward matches and assigns a unit cost to mismatches and gaps.

In considering the degree of similarity that we expect by chance, these scoring functions allow us to compare two alignments by comparing their scores, but are less useful for assessing a pairwise alignment in an absolute sense. Given a pair of aligned sequences with a particular collection of matches, mismatches, and indels, does the alignment reflect enough similarity to suggest that it is of biological interest? One way of assessing an alignment in an absolute sense is to determine whether it reflects more similarity than we would expect by chance. In developing this approach, we must take into account the divergence of related sequences due to mutation. With that in mind, we will explore models of sequence evolution and then discuss how they are used to assess alignments. Sequence evolution models are typically based on Markov chains, so we will begin with a general introduction to Markov models.

Finite discrete Markov chains

In various computational biology applications, it is useful to track the stochastic variation of a random variable. Here are some examples:

1. For models of sequence evolving by point mutation, the random variable of interest is the nucleotide observed at a fixed position, or *site*, in the sequence at time t. The goal is to characterize how this random variable changes over time.

2. It is also useful consider how the residues in a sequence change as one moves along the sequence from one site to the next. In this case, the random variable is the amino acid (or nucleotide) at site i. We are interested in how the probability of observing a given amino acid at site i depends on the amino acid observed at site $i-1$.

For each of these examples, we can model how the random variable (the nucleotide or amino acid) changes with respect to an independent variable (time or position), using a *Markov* chain with a finite number of *states*, $E_1, E_2, \ldots E_s$. Each state corresponds to one of the possible values of the random variable. In our examples, the states are defined as follows:

1. There are four states, each corresponding to the event of observing one of the four nucleotides at the site of interest, e.g., $E_1 = A, E_2 = C, E_3 = G, E_4 = T$. In a time-dependent system, such as this one, we say the system is in state E_j at time t.

2. Each of the 20 states corresponds to the event of observing a given amino acid; for example $E_1 = \text{Ala}, E_2 = \text{Cys}, \ldots E_{20} = \text{Tyr}$. In a spatially varying system, we say the system is in state E_j at site i. This is in contrast to the previous example, where time varies and the position, i, is held fixed.

The probability that a Markov chain is in state E_j at time t is designated $\varphi_j(t)$[1]. The vector $\varphi(t) = (\varphi_1(t), \varphi_2(t), \ldots \varphi_s(t))$ describes the *state probability distribution* over all states at time t. The *initial* state probability distribution is given by $\varphi(0)$. Note that Ewens and Grant [2] use π to denote the initial state distribution: $\pi = (\varphi_1(0), \varphi_2(0), \ldots \varphi_s(0))$.

In order to capture the stochastic variation of the system, we must also define the probability of making a transition from one state to another. The *transition probability,* P_{jk}, is the probability that the chain will be in state E_k at time $t+1$, given that it was in state E_j at the previous time step, t.

P is an $s \times s$ matrix specifying the probability of making a transition from any state to any other state. The rows of this matrix sum to one ($\sum_k P_{jk} = 1$) since the chain must be in some state at every time step. The columns do not have to add up to one, since there is no guarantee that you will arrive at a particular state, k.

The *Markov property* states that Markov chains are memoryless. In other words, the probability that the chain is in state E_j at time $t+1$, depends only on the state at time t and not on the past history of the states visited at times $t-1, t-2$...

[1] To simplify the exposition, we will focus on models where time is the independent variable. However, the framework is more general, and can be used to model variation with respect to other independent variables, such as the position in a sequence.

[2] Statistical Methods in Bioinformatics: An Introduction. W. Ewens, G. Grant. Springer 2001.

In this course, we will focus on discrete, finite, time-homogeneous Markov chains. These are models with a *finite* number of states, in which time (or space) is split into *discrete* steps. The assumption of discrete steps is quite natural for Example 2, because sequences of symbols are inherently discrete, but somewhat artificial for the sequence evolution model in Example 1, since time is continuous. Our models are *time-homogeneous*, because the transition matrix does not change over time.

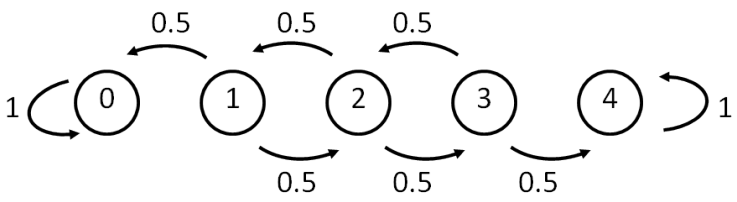
To illustrate these concepts, let us consider a simple example: A drunk is staggering about on a very short railway track with five ties on top of a mesa (a high hill with a flat top and steep sides.) At each time step, the drunk staggers either to the left or to the right. Here state E_j corresponds to the event that the drunk is standing on the j^{th} tie, where $0 \leq j \leq 4$. At each step, the drunk moves to the left or to the right with equal probability, resulting in the following transition probability matrix:

$$P = \begin{bmatrix} & E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & 1 & 0 & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Note that each row sums to one, consistent with the definition of a Markov chain.

If the drunk reaches either end of the track (either the 0^{th} or the 4^{th} tie), he falls off the mesa. This model is called a *random walk with absorbing boundaries*, because once the drunk falls off the mesa, he can never get back on the railroad track. States E_0 and E_4 are *absorbing* states. Once the system enters one of these states, it remains in that state forever, since $P_{00} = P_{44} = 1$.

The transition matrix of a Markov chain can be represented as a graph, where the nodes represent states and the edges represent transitions with non-zero probability. For example, the random walk with absorbing boundaries can be modeled like this:



How does the state probability distribution change over time? If we know the state probability distribution at time t, the distribution at the next time step is given by:

$$\varphi_k(t+1) = \sum_j \varphi_j(t) P_{jk} \tag{2}$$

or

$$\varphi(t+1) = \varphi(t) P \tag{3}$$

in matrix notation.

For example, suppose that at time $t = 0$, the drunk is standing on the middle tie; that is, $\varphi(0) = (0, 0, 1, 0, 0)$. To obtain the state probability distribution after one time step, we apply Equation 2:

$$\varphi_k(1) = \sum_{j=0}^{4} \varphi_j(0) P_{jk}.$$

For example, the probability of being in state E_1 when $t = 1$ is given by

$$\varphi_1(1) = \sum_{j=0}^{4} \varphi_j(0) P_{j1}.$$

$$= 0 \cdot 0 + 0 \cdot 0 + 1 \cdot \frac{1}{2} + 0 \cdot 0 \cdot 0 \cdot 0 \tag{4}$$

$$= \frac{1}{2}.$$

Note that Equation 4 is equivalent to multiplying the vector $(0, 0, 1, 0, 0)$ by the second column of the matrix in Equation 1.

Since the Markov chain is symmetrical, it is easy to show that $\varphi_3(1)$ is also equal to $1/2$. It is not possible to reach state E_0 or state E_4 in a single step from state E_2, so $\varphi_0(1) = \varphi_4(1) = 0$. Nor is it possible to remain in state E_2 for two consecutive time steps, so $\varphi_2(1) = 0$. Since state E_2 is the only state with non-zero probability at time $t = 0$, we obtain,

$$\varphi(1) = (0, \frac{1}{2}, 0, \frac{1}{2}, 0).$$

Now that we have the probability distribution at time $t = 1$, we can calculate the probability distribution at time $t = 2$ using the same procedure

$$\varphi_k(2) = \sum_{j=0}^{4} \varphi_j(1) P_{jk}.$$

The probability of being in state E_0 at $t = 2$ is given by

$$\begin{aligned}\varphi_0(2) &= \sum_{j=0}^{4} \varphi_j(1) P_{j0} \\ &= 0 \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} + 0 \cdot 0 + \frac{1}{2} \cdot 0 + 0 \cdot 0 \\ &= \frac{1}{4}.\end{aligned}$$

As above, $\varphi_4(2) = \varphi_0(2)$, because the matrix is symmetrical. The probability of being in state E_2 is

$$\begin{aligned}\varphi_2(2) &= \sum_{j=0}^{4} \varphi_j(1) P_{j2} \\ &= 0 \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} + 0 \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} + 0 \cdot 0 \\ &= \frac{1}{2}.\end{aligned}$$

The probabilities of being in state E_1 or E_3 at time $t = 2$ are zero, because $P_{11} = 0$ and $P_{33} = 0$. The probability distribution vector at time $t = 1$ is, therefore,

$$\varphi(2) = (\frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4}). \tag{5}$$

Summary of Markov chain notation

- A Markov chain has states $E_0, E_1, \ldots E_s$ corresponding to the range of the associated random variable.
- $\varphi_j(t)$ is the probability that the chain is in state E_j at time t. The vector $\varphi(t) = (\varphi_1(t), \ldots \varphi_s(t))$ is the state probability distribution at time t.
- $\pi = \varphi(0)$ is the initial state probability distribution.
- P is the *transition probability matrix*. P_{jk} gives the probability of making a transition to state E_k at time $t + 1$, given that the chain was in state E_j at time t. The rows of this matrix sum to one: $\sum_k P_{jk} = 1$.

- The state probability distribution at time $t+1$ is given by $\varphi(t+1) = \varphi(t) \cdot P$. The probability of being in state E_k at $t+1$ is

$$\varphi_k(t+1) = \sum_j \varphi_j(t) P_{jk}$$

- The *Markov property* states that Markov chains are memoryless. The probability that the chain is in state E_k at time $t+1$, depends only on $\varphi(t)$ and is independent of $\varphi(t-1), \varphi(t-2), \varphi(t-3) \ldots$

In this course, we will focus on discrete, finite, time-homogeneous Markov chains. These are models with a finite number of states, in which the independent variable takes on a discrete set of values. In other words, assume that time (or space) is split into discrete steps. A Markov chain is time-homogeneous if the transition matrix does not change over time.

Higher order Markov chains

Suppose that we wish to know the state of the system after two time steps. In the previous section, we used Equation 2 to calculate $\varphi(1)$, given $\varphi(0)$, and then we applied Equation 2 again to calculate $\varphi(2)$, from $\varphi(1)$. Here we derive a general expression for $\varphi(t+2)$ in terms of $\varphi(t)$. From Equation 2, we obtain

$$\varphi_l(t+1) = \sum_{j=0}^{s} \varphi_j(t) P_{jl} \tag{6}$$

and

$$\varphi_k(t+2) = \sum_{l=0}^{s} \varphi_l(t+1) P_{lk}. \tag{7}$$

Substituting the right hand side of Equation 6 for $\varphi_l(t+1)$ in Equation 7 yields

$$\varphi_k(t+2) = \sum_{l=0}^{s} \left(\sum_{j=0}^{s} \varphi_j(t) P_{jl} \right) P_{lk}.$$

We can reverse the order of the summations since the terms may be added in any order, yielding

$$\varphi_k(t+2) = \sum_{j=0}^{s} \varphi_j(t) \left(\sum_{l=0}^{s} P_{jl} P_{lk} \right). \tag{8}$$

The term in the inner summation is simply the element in row j and column k of the matrix obtained by multiplying matrix P by itself. In other words,

$$P_{jk}^{(2)} = \sum_{l=0}^{s} P_{jl} P_{lk},$$

where $P^{(2)} = P \times P$, so that Equation 8 may be rewritten as

$$\varphi_k(t+2) = \sum_{j=0}^{s} \varphi_j(t) P_{jk}^{(2)}.$$

Matrix $P^{(2)}$ is the transition matrix of a 2^{nd} order Markov chain that has the same states as the 1^{st} order Markov chain described by P. A single time step in $P^{(2)}$ is equivalent to two time steps in P. Similarly, an n^{th} order Markov chain models change after n time steps with a transition probability matrix

$$P^{(n)} = \underbrace{P \times P .. \times P}_{n \times}.$$

The n^{th} order equivalent of Equation 3 is

$$\varphi(t+n) = \varphi(t) \cdot P^{(n)}.$$

As an example, let's apply this approach to our 5-state random walk with absorbing boundaries. Recall that the transition matrix for the 1^{st} order random walk is

$$P = \begin{bmatrix} & E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & 1 & 0 & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Multiplying P times itself yields the 2^{nd} order transition matrix, $P^{(2)}$:

$$P^{(2)} = \begin{bmatrix} & E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & 1 & 0 & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ E_2 & \frac{1}{4} & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ E_3 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The state probability distribution at $t = 2$ can be calculated by applying $P^{(2)}$ to $\varphi(0)$:

$$\begin{aligned}\varphi(2) &= \varphi(0) \cdot P^{(2)} \\ &= (0, 0, 1, 0, 0) \cdot P^{(2)} \\ &= (\frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4}).\end{aligned}$$

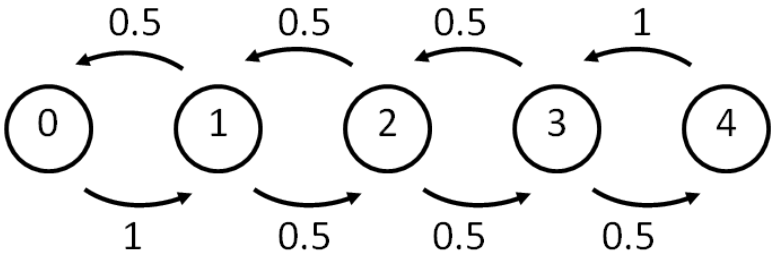
Try this matrix multiplication to convince yourself that this is correct. Note that this is the same result as Equation 5, which we got by applying the first order Markov chain twice.

Periodic Markov chains

Let us consider a second example. In order to save the drunk from an early death, we introduce a random walk with *reflecting* boundaries. At each step, the drunk moves to the left or to the right with equal probability. When the drunk reaches one of the boundary states (E_0 or E_4), he returns to the adjacent state (E_1 or E_3) at the next step, with probability one. This yields the following transition probability matrix:

$$\begin{bmatrix} & E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & 0 & 1 & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

The random walk with reflecting boundaries can be represented graphically like this:



Suppose that the drunk starts out on the middle tie at $t = 0$, as before. That is, the initial state probability distribution is $\varphi(0) = (0, 0, 1, 0, 0)$. The state distributions for the first two time steps

are the same in both random walk models, namely

$$\varphi(1) = (0, \frac{1}{2}, 0, \frac{1}{2}, 0)$$
$$\varphi(2) = (\frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4}).$$

This makes sense because the two random walk models differ only in the boundary states, E_0 and E_4, and $\varphi_0(t) = \varphi_4(t) = 0$ when $t = 0$ or $t = 1$. We calculate the state probability distribution at $t = 3$ by multiplying the vector $\varphi(2)$ with the matrix P:

$$\begin{aligned}\varphi(3) &= \varphi(2) \cdot P \\ &= (\frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4}) \cdot P \\ &= (0, \frac{1}{2}, 0, \frac{1}{2}, 0).\end{aligned}$$

This demonstrates that the state probability distribution at time $t = 3$ is the same as the distribution at time $t = 1$. In other words, $\varphi(3) = \varphi(1)$. Similarly, $\varphi(4) = \varphi(2)$, as can be seen from the following calculation:

$$\begin{aligned}\varphi(4) &= \varphi(3) \cdot P \\ &= (0, \frac{1}{2}, 0, \frac{1}{2}, 0) \cdot P \\ &= (\frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4}).\end{aligned}$$

From this we can see that the probability state distribution will be $(0, \frac{1}{2}, 0, \frac{1}{2}, 0)$ at all odd time steps and $(\frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4})$ at all even time steps. Thus, the random walk with reflecting boundaries is a periodic Markov chain. A Markov chain is *periodic* if there is some state that can only be visited in multiples of m time steps, where $m > 1$.

We do not require periodic Markov chains for modeling sequence evolution and will only consider aperiodic Markov chains going forward.

Stationary distributions

A state probability distribution, φ^*, that satisfies the equation

$$\varphi^* = \varphi^* P \tag{9}$$

is called a *stationary* distribution. A key question for a given Markov chain is whether such a stationary distribution exists. Equation 9 is equivalent to a system of s equations in s unknowns. One way to determine the steady state distribution is to solve that system of equations. The

stationary distribution can also be obtained using matrix algebra, but that approach is beyond the scope of this course.

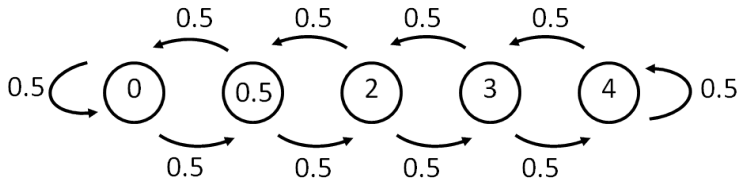
The random walk with reflecting boundaries clearly does not have a stationary distribution, since every state with non-zero probability at time t has zero probability at time $t+1$. The random walk with absorbing boundaries does have a stationary distribution, but it is not unique. Both $(1,0,0,0,0)$ and $(0,0,0,0,1)$ are stationary distributions of the random walk with absorbing boundaries.

For the rest of this course, we will concern ourselves only with aperiodic Markov chains that do not have absorbing states. In fact, we will make an even stronger assumption and restrict our consideration to Markov chains in which every state is connected to every other state via a series of zero or more states. If a finite Markov chain is aperiodic and connected in this way, it has a unique stationary distribution. We will not attempt to prove this or even to state the theorem in a rigorous way. That is beyond the scope of this class. For those who are interested, a very nice treatment can be found in Chapter 15 of *Probability Theory and its Applications (Volume I)* by William Feller (John Wiley & Sons).

As an example of a Markov chain with a unique stationary distribution, we introduce a third random walk model that has neither absorbing, nor reflecting boundaries. In this model, if the drunk is in one of the boundary states (E_0 or E_4) at time t, then at time $t+1$ he remains in the boundary state with a probability of 0.5 or returns to the adjacent state (E_1 or E_3) with a probability of 0.5. This results in the following state transition matrix

$$P = \begin{bmatrix} & E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

which can be represented graphically like this:



We can determine the stationary state distribution for this random walk model by substituting this transition matrix into Equation 2. The probability of being in state E_0 is

$$\begin{aligned}\varphi_0^* &= \sum_{j=0}^{4} \varphi_j^* P_{j0} \\ &= \varphi_0^* P_{00} + \varphi_1^* P_{10} + \varphi_2^* P_{20} + \varphi_3^* P_{30} + \varphi_4^* P_{40}.\end{aligned}$$

This reduces to

$$\varphi_0^* = \frac{1}{2}\varphi_0^* + \frac{1}{2}\varphi_1^*,$$

since P_{20}, P_{30} and P_{40} are all equal to zero. The other steady state probabilities are derived similarly, yielding

$$\varphi_0^* = \frac{1}{2}\varphi_0^* + \frac{1}{2}\varphi_1^* \tag{10}$$

$$\varphi_1^* = \frac{1}{2}\varphi_0^* + \frac{1}{2}\varphi_2^* \tag{11}$$

$$\varphi_2^* = \frac{1}{2}\varphi_1^* + \frac{1}{2}\varphi_3^* \tag{12}$$

$$\varphi_3^* = \frac{1}{2}\varphi_2^* + \frac{1}{2}\varphi_4^* \tag{13}$$

$$\varphi_4^* = \frac{1}{2}\varphi_3^* + \frac{1}{2}\varphi_4^*. \tag{14}$$

In addition, the steady state probabilities must sum to 1, imposing an additional constraint:

$$\varphi_0^* + \varphi_1^* + \varphi_2^* + \varphi_3^* + \varphi_4^* = 1. \tag{15}$$

The model has a stationary distribution if the above equations have a solution. By repeated substitution, it is possible to show that Equations 10 - 14 reduce to $\varphi_0^* = \varphi_1^* = \varphi_2^* = \varphi_3^* = \varphi_4^*$. (Do the algebra to convince yourself that this is true.) Applying the constraint in Equation 15, we see that $\varphi^* = (0.2, 0.2, 0.2, 0.2, 0.2)$ is a unique solution to the above equations.

In this example, we found a unique solution to Equation 15, demonstrating that our third random walk has a unique stationary state. Solving Equation 9 is a general approach to finding the stationary distribution. Alternatively, if we know the stationary state distribution, or have an educated guess, it is sufficient to verify that it indeed satisfies Equation 9. For example, it is easy to verify that $(0.2, 0.2, 0.2, 0.2, 0.2) \cdot P = (0.2, 0.2, 0.2, 0.2, 0.2)$.

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 and the transition probability, P_{jk}, is the probability that nucleotide j will be replaced by nucleotide k in one time stop. 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 state 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.

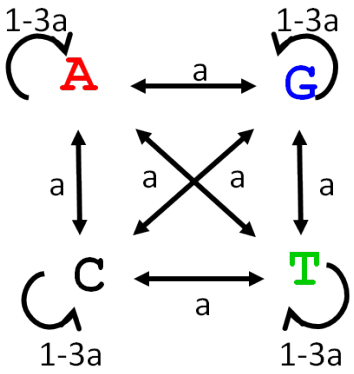
The simplest Markov model of sequence evolution for DNA is the *Jukes-Cantor model*[3], which assumes that all substitutions ($A \to C, A \to G, A \to T, C \to A...$) are equally probable and occur at a rate, α. The consequence of this assumption that the overall rate of substitution is $\lambda = 3\alpha$. That is, λ 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 $1 - 3\alpha$.

The transition probability matrix for this Markov model is:

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

[3]Jukes and Cantor, Evolution of protein molecules. In H. N. Munro, (ed.) *Mammalian Protein Metabolism*, 21-123, Adademic Press, NY, 1969.

This model can be represented graphically like this:



The stationary distribution of this Markov chain is $\varphi^* = (0.25, 0.25, 0.25, 0.25)$. (Verify that this is so using Equation 9).

The rate, α, of each possible substitution is an explicit parameter of the Jukes-Cantor model. 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 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. 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.

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 α is the rate of transitions and β is the rate of transversions.

Non-uniform stationary distributions: 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. However, this is not a realistic model for the many genomes in which the G+C content deviates from 50%. The Felsenstein (1981) model, like the Jukes-Cantor model, assumes that all substitutions are equally likely, but can model an arbitrary stationary distribution, $\varphi = (\varphi_A, \varphi_C, \varphi_G, \varphi_T)$, where $\varphi_A \neq \varphi_C \neq \varphi_G \neq \varphi_T$. The *Felsenstein model* transition matrix is

$$\begin{bmatrix} & A & G & C & T \\ A & 1-\alpha\cdot(\varphi_C+\varphi_G+\varphi_T) & \varphi_G\cdot\alpha & \varphi_C\cdot\alpha & \varphi_T\cdot\alpha \\ G & \varphi_A\cdot\alpha & 1-\alpha\cdot(\varphi_A+\varphi_C+\varphi_T) & \varphi_C\cdot\alpha & \varphi_T\cdot\alpha \\ C & \varphi_A\cdot\alpha & \varphi_G\cdot\alpha & 1-\alpha\cdot(\varphi_A+\varphi_G+\varphi_T) & \varphi_T\cdot\alpha \\ T & \varphi_A\cdot\alpha & \varphi_G\cdot\alpha & \varphi_C\cdot\alpha & 1-\alpha\cdot(\varphi_A+\varphi_C+\varphi_G) \end{bmatrix}$$

The *Hasegawa, Kishino, Yano (HKY) model*, which combines both innovations, allows different rates for transitions and transversions and an arbitrary stationary distribution, $\varphi^* = (\varphi_A, \varphi_C, \varphi_G, \varphi_T)$.

$$\begin{bmatrix} & A & G & C & T \\ A & 1-\alpha\varphi_G-\beta\cdot(\varphi_C+\varphi_T) & \varphi_G\cdot\alpha & \varphi_C\cdot\beta & \varphi_T\cdot\beta \\ G & \varphi_A\cdot\alpha & 1-\alpha\varphi_A-\beta\cdot(\varphi_C+\varphi_T) & \varphi_C\cdot\beta & \varphi_T\cdot\beta \\ C & \varphi_A\cdot\beta & \varphi_G\cdot\beta & 1-\alpha\varphi_T-\beta\cdot(\varphi_A+\varphi_G) & \varphi_T\cdot\alpha \\ T & \varphi_A\cdot\beta & \varphi_G\cdot\beta & \varphi_C\cdot\alpha & 1-\alpha\varphi_C-\beta\cdot(\varphi_A+\varphi_G) \end{bmatrix}$$

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. These models are discussed in greater detail in various molecular evolution textbooks; see, for example, Li's *Molecular Evolution*, (Sinauer Associates, 1997).

In deciding which model to use for a particular data set, we face the usual tradeoff: more general models with more parameters provide a more accurate representation of the underlying evolutionary process. However, more complex models require more data to infer 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 modeled 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 α_r. Ziheng Yang discusses this approach in his textbook *Computational Molecular Evolution* (Oxford University Press, 2006).

In addition, 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.

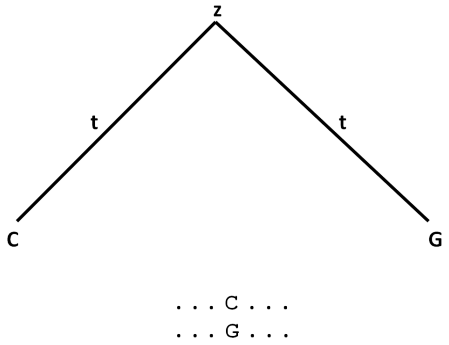
Two applications of DNA substitution models

Markov models of sequence substitution are used to answer a wide range of questions that arise in molecular evolution, including simulating sequence evolution, estimating rates of evolution, and deriving substitution scoring matrices. 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.

The likelihood of a pair of aligned sequences 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 highest. Given a parameterized model, the parameter values are estimated by determining the values that maximize the probability of the data.

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 MSA.

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.). The probability of observing x aligned with y is $P_{zx}(t) \cdot P_{zy}(t)$, the product of the probability of observing an x after time t and 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:

$$\sum_{z \in \{A,C,G,T\}} P_z P_{zx}(t) \cdot P_{zy}(t). \tag{16}$$

For example, the likelihood of observing a cytosine in one sequence and a guanine in the other is

$$\mathcal{L}(GC|t) = P_A P_{AC}(t) \cdot P_{AG}(t) + P_C P_{CC}(t) P_{CG}(t) + P_G P_{GC}(t) P_{GG}(t) + P_T P_{TC}(t) P_{TG}(t) \tag{17}$$

P_z, the frequency of z in the ancestral genome, is also unknown, but can be approximated by $\varphi^*(z)$, the frequency of z in the stationary distribution in the model.

We now derive expressions for the probability $P_{zx}(t)$ under the assumption that the sequence is evolving according to the Jukes Cantor model. The Jukes Cantor transition probability matrix gives the probability of a substitution in a single time step.

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

From this, we derive an expression describing how changes accumulate at site i over a period of time t. The probability of observing, for example, an A at site i after one time step has elapsed is given by

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

where $\varphi_j(t)$ is the probability of being in state E_j at time t. This reduces to

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

Here, the first term gives the probability that the residue at site i at time t was an A and substitution occurred. The second term is the probability that the residue at time t was not an A and a substitution occurred, 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

$$\varphi_x(t+1) = (1-4\alpha)\varphi_x(t) + \alpha.$$

Subtracting $\varphi_x(t)$ from both sides and some algebraic manipulation yields

$$\varphi_x(t+1) - \varphi_x(t) = \alpha\left(1 - 4\varphi_x(t)\right).$$

Applying a continuous time approximation allows us to express this as a differential equation

$$\frac{d\varphi_x(t)}{dt} = \alpha\left(1 - 4\varphi_x(t)\right)$$

with solution

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

We now have an expression for the probability of observing nucleotide x at site i after an arbitrarily long elapsed time, t that depends on the initial state probability $\varphi(0)_x$. We have two cases. The probability that the present-day residue is the same as the ancestral nucleotide ($\varphi_{z=x}(0) = 1$) after time t is

$$p_{xx}(t) = \frac{1}{4} \; + \; \frac{3}{4} e^{\,-4\alpha t}. \tag{18}$$

The probability that the present-day nucleotide differs from the ancestral residue ($\varphi_y(0) = 0$) is

$$p_{zx}(t) = \frac{1}{4} \; - \; \frac{1}{4} e^{\,-4\alpha t}. \tag{19}$$

Equations 18 and 19 correspond to the terms in the expression for the likelihood in Equation 16. Substituting the right hand sides of Equations 18 and 19 into Equation 17, we obtain the likelihood for observing G aligned with C

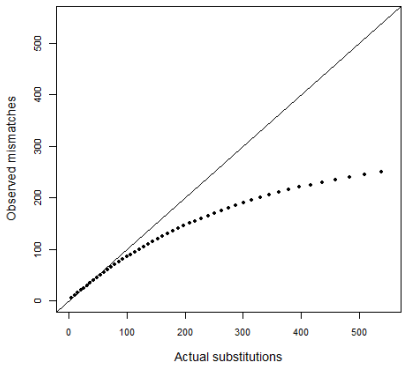
$$\mathcal{L}(G, C|t, \alpha) = \frac{1}{2}\left(\frac{1}{4} \; - \; \frac{1}{4} e^{\,-4\alpha t}\right)^2 + \frac{1}{2}\left(\frac{1}{4} \; - \; \frac{1}{4} e^{\,-4\alpha t}\right) \cdot \left(\frac{1}{4} \; + \; \frac{3}{4} e^{\,-4\alpha t}\right),$$

assuming that $P_z = \frac{1}{4}$ for all z. We now have an expression for the probability of observing C aligned with G that depends on two parameters: the branch length, t, and the substitution rate, α. These parameter values are then estimated by finding the values of t and α that maximize $\mathcal{L}(G, C|t, \alpha)$.

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 16. Under the assumption of positional independence, the likelihood for multiple sites is simply the product of the likelihoods for each site, individually.

Correcting for multiple substitutions. Another task that arises in molecular evolution is estimating the amount of sequence divergence between a pair of sequences. For example, the progressive alignment heuristic for multiple sequence alignment requires a matrix of the pairwise distances between all pairs of sequences, s_a and s_b.

A simple approach would be to count the number of positions that are not identical in the pairwise alignment of s_a and s_b. 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 distance as shown below: Recall that the Jukes-



Cantor model assumes that all substitutions ($A \to C, A \to G, A \to T, C \to A...$) are equally likely and occur at a rate α. The consequence of this assumption that the overall rate of substitution is $\lambda = 3\alpha$. Suppose that we have an ungapped[4] pairwise alignment of length n of two nucleotide sequences, σ and τ 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.

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, λ, in both lineages then the expected number of substitutions per site is

$$\begin{aligned} P_s &= 2\lambda t \\ &= 6\alpha t. \end{aligned}$$

Since both α 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} and P_{xy} that we derived in the previous section, we estimate the frequency of mismatches as a function of αt,

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

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 at

[4]None of the substitution models we have discussed account for insertions and deletions.

a single site in terms of the number of mismatches.

$$6\alpha t = f^{-1}(P_m).$$

The true frequency of mismatches is unknown, but can be estimated by $\frac{m}{n}$, yielding an equation of the form

$$6\alpha t \approx f^{-1}(\tfrac{m}{n}).$$

Finally, we multiply by n to obtain an estimate of the expected number of substitutions:

$$E[\text{sub}] = P_s n \approx f^{-1}(\tfrac{m}{n})n$$

.

First, we derive an expression for the frequency of mismatches, by 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 A at site i is

$$P_M = [p_{AA}(t)]^2 + [p_{TA}(t)]^2 + [p_{CA}(t)]^2 + [p_{GA}(t)]^2,$$

where t is the elapsed time since their divergence. Replacing the first term with Equation 18 and the remaining terms with Equation 19, 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 A's in both sequences if the ancestral nucleotide was also A. The second term represents the case where the ancestral nucleotide was not an A. By expanding the squared quantities and combining terms, we obtain

$$P_M = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}. \tag{20}$$

Note that since the Jukes Cantor model is symmetric, Equation 20 in fact gives the probability of observing the same nucleotide x in both sequences, where x may be any nucleotide. P_m, the probability of observing a mismatch at site i, is simply $1 - P_M$ or

$$P_m = \frac{3}{4}(1 - e^{-8\alpha t}). \tag{21}$$

Recall that our ultimate goal is to estimate the expected number of substitutions that actually occurred at site i since the sequences diverged. This quantity is $E[sub] = 2\lambda t \cdot n = 6\alpha t \cdot n$. We solve the above equation to obtain an expression for αt 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 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:

$$E[sub] \approx -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{m}{n}\right) \cdot n. \tag{22}$$

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

Equations that are analagous to Equations 18, 19, and 22 can be derived for the Kimura 2 Parameter (K2P) model, which assumes that transitions and transversions occur at different rates. We simply state these here with out deriving them. The probability of observing the same nucleotide after elapsed time t in the K2P model is

$$p_{xx}(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t},$$

where α is the rate of transitions and β is the rate of transversions. This equation is analogous to Equation 18. The probability of observing a different nucleotide is

$$p^s_{xy}(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t}$$

if $x \to y$ is a transition and

$$p^v_{xy} = \frac{1}{4} - \frac{1}{4}e^{-4\beta t}$$

if $x \to y$ is a transversion. Given an alignment of length n, with m_s transitions and m_v transversions, the expected number of actual substitutions is

$$E[sub] = \left[-\frac{1}{2}\ln\left(1 - \frac{2m_s}{n} - \frac{m_v}{n}\right) - \frac{1}{4}\ln\left(1 - \frac{2m_v}{n}\right)\right] \cdot n.$$