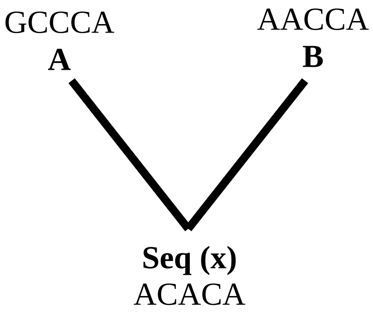
PhyloMath Lecture 2

Jonathan Richmond (03 February 2004)

Simulation of two sequences under the JC model

The main goal for the first half of the class was to simulate two sequences (A and B; the sequences we actually simulated are shown in the figure) from an ancestral sequence (x), each with 5 sites. Our first order of business was to determine the ancestral sequence (x). Remember that the JC model specifies equal base frequencies, so we can assign each base a relative frequency of 0.25 then construct the same table we used at the beginning of the last lecture (e.g. π_A has the range 0.00 to 0.25, π_C has the range 0.25 to 0.50, etc.; the format is the same as the table at the bottom of this page). After drawing random numbers, the table can be used to decide which base to insert into the simulated sequence (x). The actual sequence we generated is shown in the figure to the left.



Once we had the ancestral sequence (x), we could then simulate seq (A) and (B) separated by time (t) with a substitution rate (α). To do this, we calculated the transition probabilities for each nucleotide based on the following formulas:

$$P_{AA}(t) = ¼ + ¾ e^{-4\alpha t}$$
$$P_{AC}(t) = ¼ - ¼ e^{-4\alpha t}$$

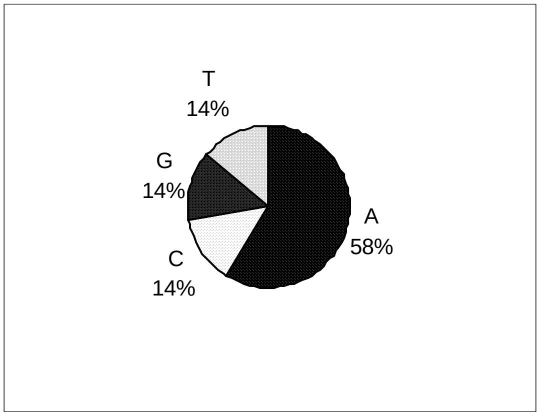
The first equation calculates the probability of ending up with nucleotide A at a particular site given that we started with A, whereas the second equation represents the probability of ending up with base C. The second equation is also used to calculate the probability of ending up with a G or T as well, since the JC model assumes that all types of substitutions are equally probable. Paul demonstrated a little shortcut for calculating the probability of a site ending up with a different base that it started with. Using the first formula above (assume $\alpha t = 0.2$), we showed that:

$$P_{AA}(t) = .587$$

Since the probabilities for the remaining types of substitutions have to equal 1 - .587,

$$P_{ij}(t) = 0.138$$

In other words, the probability of changing from base A to C, G, or T at the first site is equal to 0.138. We demonstrated this visually using a pie diagram, which is another helpful way of showing that the probability of a site ending up with the same base in the descendant sequence is greater than showing up with a different base (i.e. the biggest slice of the pie goes to staying the same). Once we had these probabilities, we then simulated our sequences (A) and (B) on a site-by-site basis using these transition probabilities. For example, we started with the first site in seq (x), which happens to be an A, and set up a table for deciding how to process the uniform random numbers.



Base	Range	Proportion of Pie
A	0.000 – 0.587	0.587
C	0.587 – 0.725	0.138
G	0.725 – 0.863	0.138
T	0.863 – 1.000	0.138

The intervals for C, G, and T are the result of sequentially adding the 0.138 transition probability until we finally get up to 1.0 for T. We could have rearranged C, G, or T for any of the ranges above 0.587 since they all have an equal transition probability under the JC model.

We can continue this exact same process for the second site, which happens to be C in the ancestral sequence (x). Going back to our original formula, $P_{CC}(t) = ¼ + ¾\ e^{-4\alpha t}$, and therefore $P_{CC}(t) = .587$. The table then turns into something like this:

Base	Range
C	0.0 – 0.587
A	0.587 – 0.725
G	0.725 – 0.863
T	0.863 – 1.0

I just switched A and the C around in the table so that $P_{Ci}(t) = 0.138$ (i.e. the probability of changing from a C to any one of the other 3 bases). In pie diagram terms, C now represents the biggest slice of pie. We iterated this process 5 times for each of the descendant sequences (A) and (B) to come up with the simulated sequences shown in the figure at the top of page 1. All this pie talk is making me hungry.

Calculating the likelihood of two sequences using the JC model

We then wanted to calculate the probability of observing sequences (A) and (B) given our model and some rate of substitution (α) over time (*t*). To do this, we had to calculate the likelihoods for each site individually and then sum the likelihoods over all sites. The way this is done is actually really straightforward, although it looks a bit daunting with all of the algebra. We can break it down so that the calculation looks something like this for site 1 (again, this all refers back to the first figure on page 1).

["L_1 =" preceded expression below in original, but this is only part of the site likelihood]

$$\Pr(\text{G in seq. A, A in seq. B} \mid \alpha, t, \text{A})$$

This states that the probability of the data for site 1 given state A in the ancestor is equal to the probability of observing a G in the descendant sequence (A), an A in the descendant sequence (B), given some rate of substitution (α) over time (*t*), and starting with an A at site 1 in the ancestral sequence. This is just one possibility for the type of mutation that can occur at this site, and since we have not observed the base in the ancestral sequence (x), we need to consider all of the possible substitution types for this site. Remember that one of the virtues of likelihood is that it is an equal opportunity employer; it takes all possible changes into account, even if they are unlikely to occur. Therefore, if we continue on for site 1 we end up with something like this for the site likelihood:

$$L_1 = \Pr(A)\ P(\text{data} \mid \alpha, t, A) + \Pr(C)\ P(\text{data} \mid \alpha, t, C) + \Pr(G)\ P(\text{data} \mid \alpha, t, G) + \Pr(T)\ P(\text{data} \mid \alpha, t, T)$$

Since we're still using the JC model, each of the probabilities Pr(A), Pr(C), Pr(G), and Pr(T) is equal to 0.25. The equation above can thus be written as:

$$L_1 = (1/4)\ P_{AG}(t)\ P_{AA}(t) + (1/4)\ P_{CG}(t)\ P_{CA}(t) + (1/4)\ P_{GG}(t)\ P_{GA}(t) + (1/4)\ P_{TG}(t)\ P_{TA}(t)$$

Next, we referred back to our handy transition equations to help us calculate these probabilities. When the site undergoes a substitution, we have:

$$\begin{aligned} P_{ij}(t) &= ¼ - ¼e^{-4\alpha t} \\ &= ¼\ (1 - e^{-4\alpha t}) \\ &= ¼\ (1 - \varnothing) \end{aligned}$$

And when the site stays the same:

$$P_{ii}(t) = \tfrac{1}{4} + \tfrac{3}{4}e^{-4\alpha t}$$
$$= \tfrac{1}{4}\,(1 + 3\varnothing)$$

We used $\varnothing = e^{-4\alpha t}$ just to simplify the notation. Substituting these values:

$$L_1 = (1/4)\ P_{AG}(t)\ P_{AA}(t) + (1/4)\ P_{CG}(t)\ P_{CA}(t) + (1/4)\ P_{GG}(t)\ P_{GA}(t) + (1/4)\ P_{TG}(t)\ P_{TA}(t)$$

$$\begin{gathered} L_1 = (\tfrac{1}{4})\ [(\tfrac{1}{4})(1 - \varnothing)][(\tfrac{1}{4})\ (1 + 3\varnothing)] \\ + (\tfrac{1}{4})\ [(\tfrac{1}{4})(1 - \varnothing)][(\tfrac{1}{4})\ (1 - \varnothing)] \\ + (\tfrac{1}{4})\ [(\tfrac{1}{4})(1 + 3\varnothing)][(\tfrac{1}{4})\ (1 - \varnothing)] \\ + (\tfrac{1}{4})\ [(\tfrac{1}{4})(1 - \varnothing)][(\tfrac{1}{4})\ (1 - \varnothing)] \end{gathered}$$

We then simplified the equation L_1 above as follows:

$L_1 = (\tfrac{1}{4})\ [(\tfrac{1}{4})(1 - \varnothing)][(\tfrac{1}{4})\ (1 + 3\varnothing)] + (\tfrac{1}{4})\ [(\tfrac{1}{4})(1 - \varnothing)]^2 + (\tfrac{1}{4})\ [(\tfrac{1}{4})(1 - \varnothing)][(\tfrac{1}{4})\ (1 + 3\varnothing)] + (\tfrac{1}{4})\ [(\tfrac{1}{4})(1 - \varnothing)]^2$

$L_1 = (\tfrac{1}{4})^3(1 - \varnothing)(1 + 3\varnothing) + (\tfrac{1}{4})^3(1 - \varnothing)^2 + (\tfrac{1}{4})^3(1 - \varnothing)(1 + 3\varnothing) + (\tfrac{1}{4})^3(1 - \varnothing)^2$

$L_1 = 2\ (\tfrac{1}{4})^3(1 - \varnothing)(1 + 3\varnothing) + 2(\tfrac{1}{4})^3(1 - \varnothing)^2$

$L_1 = 2\ (\tfrac{1}{4})^3(1 - \varnothing)[1 + 3\varnothing + 1 - \varnothing]$

$L_1 = 2\ (\tfrac{1}{4})^3(1 - \varnothing)[2 + 2\varnothing]$

$L_1 = 4\ (\tfrac{1}{4})^3(1 - \varnothing)(1 + \varnothing)$

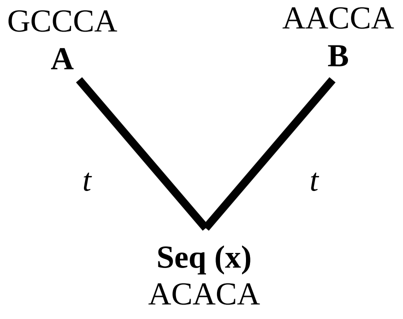
$L_1 = (\tfrac{1}{4})^2(1 - \varnothing)(1 + \varnothing)$

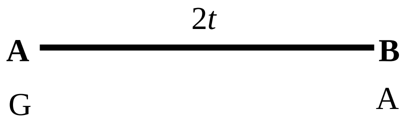
$L_1 = (\tfrac{1}{4})^2(1 - \varnothing^2)$

$L_1 = (1/16)(1 - e^{-8\alpha t})$

An important point to note is that we would get the same quantity no matter where we place the root on our tree. This is where the so-called ‘time reversibility’ of the maximum likelihood models comes into play. You can see this in the following example.

With two sequences, there is an easier way to arrive at the same value for calculating the likelihood of site 1 (i.e. $L_1 = 1/16\ (1 - e^{-8\alpha t})$). Because A and B are separated by branch length 2*t*, we can calculate the probability of changing from a G to an A using the following equation:





$$\begin{aligned} L_1 &= \Pr(G)\ P(G \rightarrow A\ \vert 2t) \\ &= (\tfrac{1}{4})\ P_{GA}(2t) \\ &= (\tfrac{1}{4})(\tfrac{1}{4})(1 - e^{-4\alpha(2t)}) \\ &= (1/16)(1 - e^{-8\alpha t}) \end{aligned}$$

Note that this is the same answer we determined algebraically above, and we got there with far less sweat this time around. I referred to (*t*) as branch length, but it might be better to think of it as time *t* since we’re assuming that the branch length is proportional to time. The purpose of this little exercise was to show that we could get the same quantity no matter where we rooted the tree. In other words, A could have been the ancestral sequence whereas B and X could have been the descendant sequences.

We just showed what went into the calculation for the likelihood of site 1, so to get the likelihood for the entire sequence we would simply multiply the likelihoods for all sites.

$$L = (L_1)(L_2)(L_3)(L_4)(L_5)$$

It's easy to see that the overall likelihood for a more normal stretch of sequence data (e.g. thousands of base pairs these days) would get small really quick unless we used log-likelihoods. You can refer back to the notes from Lecture 1 to see why this is the case. Instead, we do it like this:

$$\ln L = \ln L_1 + \ln L_2 + \ln L_3 + \ln L_4 + \ln L_5$$

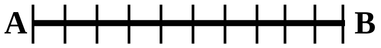
Partitioned likelihood models are rapidly gaining in popularity in modern times. The partition models work something like this:

$$\ln L = (\ln L_1 + \ln L_2 + \ln L_3) + (\ln L_4 + \ln L_5)$$

In this case, each set of bracketed likelihood scores are calculated using different models of nucleotide substitution. The likelihoods are then summed within each partition, and the cumulative scores for each partition are used to generate the overall likelihood for the entire sequence.

The Dumb Substitution Model

For the next exercise, we assumed that the sequences A and B were separated by some time interval t = 10 trillion yrs.



The probability of observing a substitution at any one of these intervals is equal to p, and the probability of not having a substitution is 1 - p. Only one substitution is allowed per interval. By setting bounds such as the maximum number of allowable substitutions over time (t) and arbitrarily choosing 10 intervals instead of 20 (or 30, 40, 100, etc.), we are potentially sacrificing some biological realism. We are really interested in calculating an expected number of substitutions over time (t) without imposing subjective limits, but this requires a somewhat complicated procedure.

Before we get to that point, we first wanted to go through a 'dumb' model, which will help us understand the more complicated model in the upcoming seminars. The dumb model involves a Bernoulli Probability Distribution, where the outcome of any single trial can be one of two states, such as failure/success or true/false. We set up a simple table to demonstrate this:

$$\text{False} = 0 = 1\text{-}p$$
$$\text{True} = 1 = p$$

Thus, $p + (1\text{-}p) = 1$. This states that the probability of something must be 1. That is, there is either a substitution or there isn't one, and there are no other possibilities.

We then performed a Bernoulli trial for each time interval, which is analogous to flipping a coin[1]. Think of heads as a success (or having a substitution in any one of those time intervals) and tails as a failure (or not having substitution). If we were actually performing the tosses, we have flipped the coin 10 times, which corresponds to the number of intervals we chose over time (t). Obviously, you can end up getting different combinations of heads and tails each time you did a set of 10 tosses. What we wanted to do was calculate the probabilities for having different numbers of successes or failures over the 10 intervals. We only care about the number of successes, not about which intervals experience the success.

[1] Note that everything boils down to a coin toss if you are Paul Lewis or colored socks if you are Kent Holsinger.

Consider the following table and assume arbitrarily that p = 0.25:

# substitutions	Probability	k Pr(k)	Cumulative
k	Pr(k)		Σ_k k pr(k)
y = 0	$(1 – p)^{10}$	0	0
y = 1	$10p(1 – p)^9$	0.1887	0.1887
y = 2	$45p^2(1 – p)^8$	0.5621	0.7508
y = 3	$120p^3(1 – p)^7$	0.7512	1.502
y = 4	$210p^4(1 – p)^6$	0.584	2.086
y = 5	$252p^5(1 – p)^5$	0.292	2.378
y = 6	$210p^6(1 – p)^4$	0.103	2.475
y = 7	$120p^7(1 – p)^3$	0.021	2.496
y = 8	$45p^8(1 – p)^2$	0.0037	2.4997
y = 9	$10p^9(1 – p)$	0.0002	2.4999
y = 10	P^{10}	0.0001	**2.5***

Starting with the 1st cell in column 1, y = 0 represents a situation where you flip only tails (i.e. not having a substitution in any of the 10 intervals) in all 10 coin tosses. The probability of that happening is represented in the 1st cell of the 2nd column, where the value (1-*p*) is raised to the 10th power. The chance that you would flip 10 heads in a row (succeeding, or having a substitution in every time interval) is represented by *y* = 10, and the probability of that happening is (p) raised to the 10th power. All values of *y* in between represent the remaining possibilities and their associated probabilities. How did we get these equations for the probabilities?

Remember that the value (1-*p*) comes from the probability of failure whereas the probability of success is just (*p*). We then needed to figure out the different **number of ways** you could get the various combinations of successes and failures. In other words, if you had 2 heads and 8 tails for your 10 coin tosses (*y* = 2 in the table above), you could have landed a heads on your first and fourth toss, or your sixth and ninth toss, or your ninth and tenth toss, etc. To do this, we use the following equation:

$$\text{no. ways of choosing y successes in n trials} = n!/y!(n-y)!$$

For *y* = 2, this works out to:

$$\begin{aligned}\text{no. ways of choosing 2 successes in 10 trials} &= 10!/(2!)(8!)\\ &= (10)(9)(8!)/(2!)(8!)\\ &= (10)(9)/2\\ &= 45\end{aligned}$$

We used this same equation to fill in all of the probabilities for each value (*y*), which is where the numbers 10, 45, 120, 210, and 252 come from. The 3rd column is associated with our calculation of the expected value of the mean number of substitutions for the 10 time intervals. We want to know that out of 10 flips of the coin, what is the mean number of times we expect to get heads (or successes, or a substitution in any one of the 10 time intervals). This is calculated using the following equation:

$$E(y) = \Sigma_k \text{ k pr(k)}$$

The above formula is the definition of the "expected value" of a random variable *y*. The column labeled "Cumulative Σ_k *k* pr(*k*)" is just the sum over all of the individual quantities in column 3, which I have shown in the fourth column. The 4th column was a little misleading the way it was drawn in class, so I added the 3rd column to try and clarify what went on. In the 3rd column, you can see that *y* = 3 makes the highest contribution, whereas *y* = 0 and *y* = 10 make very small contributions. This makes intuitive sense because the chances of flipping 10 heads or 10 tails in a row seems extremely low, whereas the chances of flipping 3 heads out of 10 tosses seems quite reasonable. When we summed each of the individual contributions in column 3, we found that the mean number of substitutions (or ‘heads’ on the coin) was expected to equal 2.5.

We ended class by imagining that we could divide our magic A-to-B time span into an infinite number of intervals. By doing this, we have moved away from explaining our data in terms of a binomial distribution to something called a Poisson distribution. Whereas the binomial distribution forces you to arbitrarily choose the number of intervals in our time sequence, the Poisson distribution allows you to evaluate an infinite number of successively smaller divisions. We are going to discuss the advantages of this approach in the upcoming lectures.