Calculating Likelihoods on Phylogenetic trees

John P. Huelsenbeck

July 24, 2012

1 Assumptions of phylogenetic methods

The models used in phylogenetic analysis of molecular data have three components. First, they assume a tree relating the samples. Here, the samples might be DNA sequences collected from different species, or different individuals within a population. In either case, a basic assumption is that the samples are related to one another through an (unknown) tree. This would be a species tree for sequences sampled from different species, or perhaps a coalescence tree for sequences sampled from individuals from within a population. Second, they assume that the branches of the tree have an (unknown) length. Ideally, the length of a branch on a tree is in terms of time. However, in practice it is difficult to determine the duration of a branch on a tree in terms of time. Instead, the lengths of the branches on the tree are in terms of expected change per character. Figure 1 shows some examples of trees with branch lengths. The main points the reader should remember are: (1) Trees can be rooted or unrooted. Rooted trees have a time direction whereas unrooted trees do not. Most methods of phylogenetic inference, including most implementations of maximum likelihood and Bayesian analysis, are based on time-reversible models of evolution that produce unrooted trees, which must be rooted using some other criterion, such as the outgroup criterion (using distantly related reference sequences to locate the root). (2) The space of possible trees is huge. The number of possible unrooted trees for n species is $B(n) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$ (Schröder, 1870). This means that for a relatively small problem of only $n = 50$ species, there are about $B(50) = 2.838 \times 10^{74}$ possible unrooted trees that can explain the phylogenetic relationships of the species.

The third component of a phylogenetic model is a process that describes how the characters change on the phylogeny. All model-based methods of phylogenetic inference, including maximum likelihood and Bayesian estimation of phylogeny, currently assume that character change occurs according to a continuous-time Markov chain. At the heart of any continuous-time Markov chain is a matrix of rates, specifying the rate of change from one state to another. For example, the instantaneous rate of change under the model

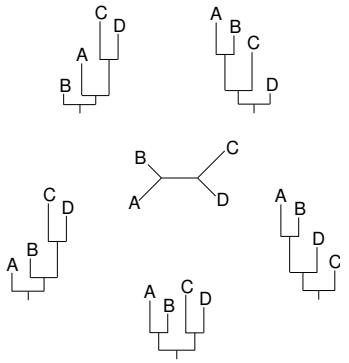

Figure 1: **Example of unrooted and rooted trees**. An unrooted tree of four species (center) with the branch lengths drawn proportional to their length in terms of expected number of substitutions per site. The five trees surrounding the central, unrooted, tree show the five possible rooted trees that result from the unrooted tree.

described by Hasegawa et al. (1984, 1985; hereafter called the HKY85 model) is

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix} \mu \quad (1)$$

This matrix specifies the rate of change from one nucleotide to another; the rows and columns of the matrix are ordered A, C, G, T, so that the rate of change from $C \rightarrow G$ is $q_{CG} = \pi_G$. Similarly, the rates of change between $C \rightarrow T$, $G \rightarrow A$, and $T \rightarrow C$, are $q_{CT} = \kappa\pi_T$, $q_{GA} = \kappa\pi_A$, and $q_{TG} = \pi_G$, respectively. The diagonals of the rate matrix, denoted with the dashes, are specified such that each row sums to zero. Finally, the rate matrix is rescaled such that the mean rate of substitution is one. This can be accomplished by setting $\mu = -1/\sum_{i\in\{A,C,G,T\}} \pi_i q_{ii}$. This rescaling of the rate matrix such that the mean rate is one allows the branch lengths on the phylogenetic tree to be interpreted as expected number of nucleotide substitutions per site.

We will make a few important points about the rate matrix. First, the rate matrix may have free parameters. For example, the HKY85 model has the parameters κ, π_A, π_C, π_G, and π_T. The parameter κ is the transition/transversion rate bias; when $\kappa = 1$ transitions occur at the same rate as transversions. Typically, the transition/transversion rate ratio, estimated using maximum likelihood or Bayesian inference, is greater than one; transitions occur at a higher rate than transversions. The other parameters—π_A, π_C, π_G, and π_T—are the base frequencies, and have a biological interpretation as the frequency of the different nucleotides and are also, incidentally, the stationary probabilities of the process (more on stationary probabilities later). Second, the rate matrix, $\mathbf{Q}$, can be used to calculate the transition probabilities and the stationary distribution of the substitution process. The transition probabilities and stationary distribution play a key role in calculating the likelihood, and we will spend more time here developing an intuitive understanding of these concepts.

1.1 Transition probabilities

Let us consider a specific example of a rate matrix, with all of the parameters of the model taking specific values. For example, if we use the HKY85 model and fix the parameters to $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$, we get the following matrix of instantaneous rates

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

Note that these numbers are not special in any particular way. That is to say, they are not based upon any observations from a real data set, but are rather arbitrarily picked to illustrate a point. The point is that one can interpret the rate matrix in the physical sense of specifying how changes occur on a phylogenetic tree. Consider the very simple case of a single branch on a phylogenetic tree. Let's assume that the branch is $v = 0.5$ in length and that the ancestor of the branch is the nucleotide G. The situation we have is something like that shown in Figure 2A. How can we simulate the evolution of the site starting from the G at the ancestor? The rate matrix tells us how to do this. First of all, because the current state of the process is G, the only relevant row of the rate matrix is the third one:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1.266 & 0.190 & -1.519 & 0.063 \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

The overall rate of change away from nucleotide G is $q_{GA} + q_{GC} + q_{GT} = 1.266 + 0.190 + 0.063 = 1.519$. Equivalently, the rate of change away from nucleotide G is simply $-q_{GG} = 1.519$. In a continuous-time Markov model, the waiting time between substitutions is exponentially distributed. The exact shape of the exponential distribution is determined by its rate, which is the same as the rate of the corresponding process

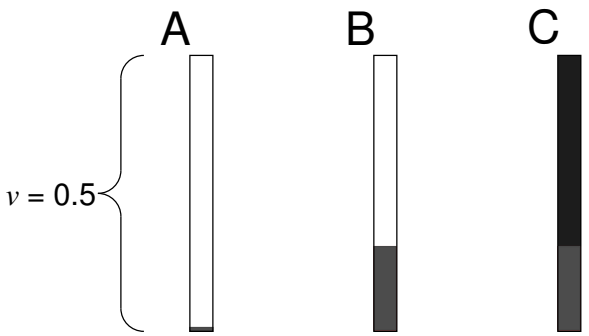

Figure 2: **Simulation under the HKY85 substitution process**. A single realization of the substitution process under the HKY85 model when $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$. The length of the branch is $v = 0.5$ and the starting nucleotide is G (light gray). A, The process starts in nucleotide G. B, The first change is 0.152 units up the branch. C, the change is from G to A (dark gray). The time at which the next change occurs exceeds the total branch length, so the process ends in state C.

---

in the **Q** matrix. For instance, if we are in state G, we wait an exponentially distributed amount of time with rate 1.519 until the next substitution occurs. One can easily construct exponential random variables from uniform random variables using the equation

$$t = -\frac{1}{\lambda}\log_e(u)$$

where λ is the rate and u is a uniform(0,1) random number. For example, my calculator has a uniform(0,1) random number generator. The first number it generated is $u = 0.794$. This means that the next time at which a substitution occurs is 0.152 up from the root of the tree (using $\lambda = 1.519$; Figure 2B). The rate matrix also specifies the probabilities of a change from G to the nucleotides A, C, and T. These probabilities are

$$G \to A : \tfrac{1.266}{1.519} = 0.833, \quad G \to C : \tfrac{0.190}{1.519} = 0.125, \quad G \to T : \tfrac{0.063}{1.519} = 0.042$$

To determine what nucleotide the process changes to we would generate another uniform(0,1) random number (again called u). If u is between 0 and 0.833, we will say that we had a change from G to A. If the random number is between 0.833 and 0.958 we will say that we had a change from G to C. Finally, if the random number u is between 0.958 and 1.000, we will say we had a change from G to T. The next number generated on our calculator was $u = 0.102$, which means the change was from G to A. The process is now in a different state (the nucleotide A) and the relevant row of the rate matrix is

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

We wait an exponentially distributed amount of time with parameter $\lambda = 0.886$ until the next substitution occurs. When the substitution occurs, it is to a C, G, or T with probabilities $\frac{0.190}{0.886} = 0.214$, $\frac{0.633}{0.886} = 0.714$, and $\frac{0.063}{0.886} = 0.072$, respectively. This process of generating random and exponentially distributed times until the next substitution occurs and then determining (randomly) what nucleotide the change is to is repeated until the process exceeds the length of the branch. The state the process is in when it passes the end of the branch is recorded. In the example of Figure 2, the process started in state G and ended in state A. (The next uniform random variable generated on our calculator was $u = 0.371$, which means that the next substitution would occur 1.119 units above the substitution from $G \to A$. The process is in the state A when it passed the end of the branch.) The only non-random part of the entire procedure was the initial decision to start the process in state G. All other aspects of the simulation used a uniform random number generator and our knowledge of the rate matrix to simulate a single realization of the HKY85 process of DNA substitution.

This Monte Carlo procedure for simulating the HKY85 process of DNA substitution can be repeated. The following table summarizes the results of 100 simulations, each of which started with the nucleotide G:

Starting Nucleotide	Ending Nucleotide	Number of Replicates
G	A	27
G	C	10
G	G	59
G	T	4

This table can be interpreted as a Monte Carlo approximation of the *transition probabilities* from nucleotide G to nucleotide $i \in (A, C, G, T)$. Specifically, the Monte Carlo approximations are $p_{GA}(0.5) \approx 0.27$, $p_{GC}(0.5) \approx 0.10$, $p_{GG}(0.5) \approx 0.59$, and $p_{GT}(0.5) \approx 0.04$. These approximate probabilities are all conditioned on the starting nucleotide being G and the branch length being $v = 0.5$. We performed additional simulations in which the starting nucleotide was A, C, or T. Together with the earlier Monte Carlo simulation that started with the nucleotide G, these additional simulations allow us to fill out the following table with the approximate transition probabilities:

		Ending Nucleotide			
		A	C	G	T
Starting Nucleotide	A	0.67	0.13	0.20	0.00
	C	0.13	0.70	0.07	0.10
	G	0.27	0.10	0.59	0.04
	T	0.12	0.30	0.08	0.50

Clearly, these numbers are only crude approximations to the true transition probabilities; after all, each row in the table is based on only 100 Monte Carlo simulations. However, they do illustrate the meaning of the transition probabilities; the transition probability, $p_{ij}(v)$, is the probability that the substitution process ends in nucleotide j conditioned on it starting in nucleotide i after an evolutionary amount of time v. The table of approximate transition probabilities, above, can be interpreted as a matrix of probabilities, usually denoted $\mathbf{P}(v)$. Fortunately, we do not need to rely on Monte Carlo simulation to approximate the transition probability matrix. Instead, we can calculate the transition probability matrix exactly using matrix exponentiation:

$$\mathbf{P}(v) = e^{\mathbf{Q}v}$$

For the case we have been simulating, the exact transition probabilities (to four decimal places) are

$$\mathbf{P}(0.5) = \{p_{ij}(0.5)\} = \begin{pmatrix} 0.7079 & 0.0813 & 0.1835 & 0.0271 \\ 0.1085 & 0.7377 & 0.0542 & 0.0995 \\ 0.3670 & 0.0813 & 0.5244 & 0.0271 \\ 0.1085 & 0.2985 & 0.0542 & 0.5387 \end{pmatrix}$$

The transition probability matrix accounts for all the possible ways the process could end up in nucleotide j after starting in nucleotide i. In fact, each of the infinite possibilities is weighted by its probability under the substitution model.

1.1.1 Stationary distribution

The transition probabilities provide the probability of ending in a particular nucleotide after some specific amount of time (or opportunity for substitution, v). These transition probabilities are conditioned on starting in a particular nucleotide. What do the transition probability matrices look like as v increases? The following transition probability matrices show the effect of increasing branch length:

$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \quad \mathbf{P}(0.01) = \begin{pmatrix} 0.991 & 0.002 & 0.006 & 0.001 \\ 0.003 & 0.993 & 0.001 & 0.003 \\ 0.013 & 0.002 & 0.985 & 0.001 \\ 0.003 & 0.009 & 0.001 & 0.987 \end{pmatrix}$$

$$\mathbf{P}(0.10) = \begin{pmatrix} 0.919 & 0.018 & 0.056 & 0.006 \\ 0.024 & 0.934 & 0.012 & 0.029 \\ 0.113 & 0.018 & 0.863 & 0.006 \\ 0.025 & 0.086 & 0.012 & 0.877 \end{pmatrix} \quad \mathbf{P}(0.50) = \begin{pmatrix} 0.708 & 0.081 & 0.184 & 0.027 \\ 0.106 & 0.738 & 0.054 & 0.100 \\ 0.367 & 0.081 & 0.524 & 0.027 \\ 0.109 & 0.299 & 0.054 & 0.539 \end{pmatrix}$$

$$\mathbf{P}(1.00) = \begin{pmatrix} 0.580 & 0.141 & 0.232 & 0.047 \\ 0.188 & 0.587 & 0.094 & 0.131 \\ 0.464 & 0.141 & 0.348 & 0.047 \\ 0.188 & 0.394 & 0.094 & 0.324 \end{pmatrix} \quad \mathbf{P}(5.00) = \begin{pmatrix} 0.411 & 0.287 & 0.206 & 0.096 \\ 0.383 & 0.319 & 0.192 & 0.106 \\ 0.411 & 0.287 & 0.206 & 0.096 \\ 0.383 & 0.319 & 0.192 & 0.107 \end{pmatrix}$$

$$\mathbf{P}(10.0) = \begin{pmatrix} 0.401 & 0.299 & 0.200 & 0.099 \\ 0.399 & 0.301 & 0.199 & 0.100 \\ 0.401 & 0.299 & 0.200 & 0.099 \\ 0.399 & 0.301 & 0.199 & 0.100 \end{pmatrix} \quad \mathbf{P}(100) = \begin{pmatrix} 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \end{pmatrix}$$

(Each matrix was calculated under the HKY85 model with $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$.) Note that as the length of a branch, v, increases, the probability of ending up in a particular nucleotide converges to a single number, regardless of the starting state. For example, the probability of ending up in C is about 0.300 when the branch length is $v = 100$. This is true regardless of whether the process starts in A, C, G, or T. The substitution process has in a sense 'forgotten' its starting state.

The stationary distribution is the probability of observing a particular state when the branch length increases without limit ($v \to \infty$). The stationary probabilities of the four nucleotides are $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$ for the example discussed above. The models typically used in phylogenetic analyses have the stationary probabilities built into the rate matrix, $\mathbf{Q}$. You will notice that the rate matrix for the HKY85 model has parameters π_A, π_C, π_G, and π_T, and that the stationary frequencies of the four nucleotides for our example match the input values for our simulations. Building the stationary frequency of the process into the rate matrix, while somewhat unusual, makes calculating the likelihood function easier. For one, specifying the stationary distribution saves the time of figuring out what the stationary distribution is (which involves solving the equation $\pi\mathbf{Q} = \mathbf{0}$, which simply says that, if we start with the nucleotide frequencies reflecting the stationary distribution, the process will have no effect on the nucleotide frequencies). For another, it allows one to more easily specify a time reversible substitution model. [A time reversible substitution model has the property that $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i, j \in (A, C, G, T)$, $i \neq j$.] Practically speaking, time reversibility means that we can work with unrooted trees instead of rooted trees (assuming that the molecular clock is not enforced).

1.1.2 Calculating the likelihood

The transition probabilities and stationary distribution are used when calculating the likelihood. For example, consider the following alignment of sequences for five species[1]:

```
Species 1   TAACTGTAAAGGACAACACTAGCAGGCCAGACGCACACGCACAGCGCACC
Species 2   TGACTTTAAAGGACGACCCTACCAGGGCGGACACAAACGGACAGCGCAGC
Species 3   CAAGTTTAGAAAACGGCACCAACACAACAGACGTATGCAACTGACGCACC
Species 4   CGAGTTCAGAAGACGGCACCAACACAGCGGACGTATGCAGACGACGCACC
Species 5   TGCCCTTAGGAGGCGGCACTAACACCGCGGACGAGTGCGGACAACGTACC
```

This is clearly a rather small alignment of sequences to use for estimating phylogeny, but it will illustrate how likelihoods are calculated. The likelihood is the probability of the alignment of sequences, conditioned on a tree with branch lengths. The basic procedure is to calculate the probability of each site (column) in the matrix. Assuming that the substitutions are independent across sites, the probability of the entire alignment is simply the product of the probabilities of the individual sites.

How is the likelihood at a single site calculated? Figure 3 shows the observations at the first site (T, T, C, C, and T) at the tips of one of the possible phylogenetic trees for five species. The tree in Figure 3 is unusual in that we will assume that the nucleotide states at the interior nodes of the tree are also known.

[1]This alignment was simulated on the tree of Figure 3 under the HKY85 model of DNA substitution. Parameter values for the simulation can be found in the caption of Table 1.

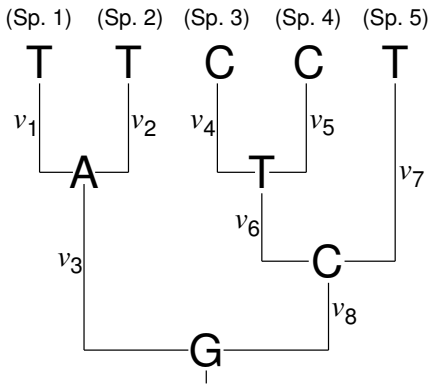

Figure 3: **A tree with states assigned to the tips**. One of the possible (rooted) trees describing the evolutionary history of the five species. The states at the first site in the alignment of the text are shown at the tips of the tree. The states at the interior nodes of the tree are also shown, though in reality these states are not observed. The length of the ith branch is denoted v_i.

This is clearly a bad assumption, because we cannot directly observe the nucleotides that occurred at any point on the tree in the distant past. For now, however, ignore this fact and bear with us. The probability of observing the configuration of nucleotides at the tips and interior nodes of the tree in Figure 3 is

$$\Pr(TTCCT, ATCG|\tau, \mathbf{v}, \theta) = \\ \pi_G\, p_{GA}(v_3)\, p_{AT}(v_1)\, p_{AT}(v_2)\, p_{GC}(v_8)\, p_{CT}(v_6)\, p_{CT}(v_7)\, p_{TC}(v_4)\, p_{TC}(v_5)$$

Here we show the probability of the observations (TTCCT) and the states at the interior nodes of the tree (ATCG) conditioned on the tree (τ), branch lengths ($\mathbf{v}$), and other model parameters (θ). Note that to calculate the probability of the states at the tips of the tree, we used the stationary probability of the process (π) and also the transition probabilities [$p_{ij}(v)$]. The stationary probability of the substitution process was used to calculate the probability of the nucleotide at the root of the tree. In this case, we are assuming that the substitution process has been running a very long time before it reached the root of our five species tree. We then use the transition probabilities to calculate the probabilities of observing the states at each end of the branches. When taking the product of the transition probabilities, we are making the additional assumption that the substitutions on each branch of the tree are independent of one another. This is probably a reasonable assumption for real data sets.

The probability of observing the states at the tips of the tree, described above, was conditioned on the interior nodes of the tree taking specific values (in this case $ATCG$). To calculate the unconditional probability of the observed states at the tips of the tree, we sum over all possible combinations of nucleotide states that can be assigned to the interior nodes of the tree

$$\Pr(TTCCT|\tau, \mathbf{v}, \theta) = \sum_w \sum_x \sum_y \sum_z \Pr(TTCCT, wxyz|\tau, \mathbf{v}, \theta)$$

where $w, x, y, z \in (A, C, G, T)$. Averaging the probabilities over all combinations of states at the interior nodes of the tree accomplishes two things. First, we remove the assumption that the states at the interior nodes take specific values. Second, because the transition probabilities account for all of the possible ways we could have state i at one end of a branch and state j at the other, the probability of the site is also averaged over all possible character histories. Here, we think of a character history as one realization of changes on the tree that is consistent with the observations at the tips of the tree. For example, the parsimony method, besides calculating the minimum number of changes on the tree, also provides a character history; the character history favored by parsimony is the one that minimizes the number of changes required to explain the data. In the case of likelihood-based methods, the likelihood accounts for all possible character histories, with each history weighted by its probability under the substitution model. Nielsen (2002) described a method for sampling character histories in proportion to their probability that relies on the

Site	Prob.	Site	Prob.	Site	Prob.	Site	Prob.	Site	Prob.
1	0.004025	11	0.029483	21	0.179392	31	0.179392	41	0.003755
2	0.001171	12	0.006853	22	0.001003	32	0.154924	42	0.005373
3	0.008008	13	0.024885	23	0.154924	33	0.007647	43	0.016449
4	0.002041	14	0.154924	24	0.179392	34	0.000936	44	0.029483
5	0.005885	15	0.007647	25	0.005719	35	0.024885	45	0.154924
6	0.000397	16	0.024124	26	0.001676	36	0.000403	46	0.047678
7	0.002802	17	0.154924	27	0.000161	37	0.024124	47	0.010442
8	0.179392	18	0.004000	28	0.154924	38	0.154924	48	0.179392
9	0.024124	19	0.154924	29	0.001171	39	0.011088	49	0.002186
10	0.024885	20	0.004025	30	0.047678	40	0.000161	50	0.154924

Table 1: **Probabilities of individual sites**. The probabilities of the fifty sites for the example alignment from the text. The likelihoods are calculated assuming the tree of Figure 3 with the branch lengths being $v_1 = 0.1$, $v_2 = 0.1$, $v_3 = 0.2$, $v_4 = 0.1$, $v_5 = 0.1$, $v_6 = 0.1$, $v_7 = 0.2$, and $v_8 = 0.1$. The substitution model parameters were also fixed, with $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$.

interpretation of the rate matrix as specifying waiting times between substitutions. His method provides a means to reconstruct the history of a character that does not inherit the flaws of the parsimony method. Namely, Nielsen's method allows multiple changes on a single branch and also allows for non-parsimonious reconstructions of a character's history.

Before moving on, we will make two final points. First, in practice no computer program actually evaluates all combinations of nucleotides that can be assigned to the interior nodes of a tree when calculating the probability of observing the data at a site. There are simply too many combinations for trees of even small size. For example, for a tree of 100 species, there are 99 interior nodes and 4.02×10^{59} combinations of nucleotides at the ancestral nodes on the tree. Instead, Felsenstein's (1981) pruning algorithm is used to calculate the likelihood at a site. Felsenstein's method is mathematically equivalent to the summation shown above, but can evaluate the likelihood at a site in a fraction of the time it would take to plow through all combinations of ancestral states. Second, the overall likelihood of a character matrix is the product of the site likelihoods. If we assume that the tree of Figure 3 is correct (with all of the parameters taking the values specified in the caption of Table 1), then the probability of observing the data is

$$0.004025 \times 0.001171 \times 0.008008 \times \ldots \times 0.154924 = 1.2316 \times 10^{-94}$$

where there are fifty factors, each factor representing the probability of an individual site (column) in the alignment. Table 1 shows the probabilities of all fifty sites for the tree of Figure 3. Note that the overall probability of observing the data is a very small number ($\approx 10^{-94}$). This is typical of phylogenetic problems and results from the simple fact that many numbers between 0 and 1 are multiplied together. Computers cannot accurately hold very small numbers in memory. Programmers avoid this problem of computer "underflow" by using the log probability of observing the data. The log probability of observing the sample alignment of sequences presented earlier is $\log_e \ell = \log_e(1.2316 \times 10^{-94}) = -216.234734$. The log likelihood can be accurately stored in computer memory.

2 Four equivalent ways to simulate DNA sequences on a tree

In the previous section, we covered some of the most basic assumptions made in a phylogenetic analysis; DNA sequences are assumed to evolve on a phylogenetic tree (with branch lengths) under a continuous-time Markov model of DNA substitution. The substitution process is assumed to be independent across sites. These assumptions apply to maximum likelihood, Bayesian inference, and distance methods (when the distances are 'corrected' under some evolutionary model). It is not so clear what assumptions the parsimony method is making (one potential criticism of the method), but several authors have found interesting connections between the parsimony method and maximum likelihood under an over-parameterized continuous-time Markov model of DNA substitution.

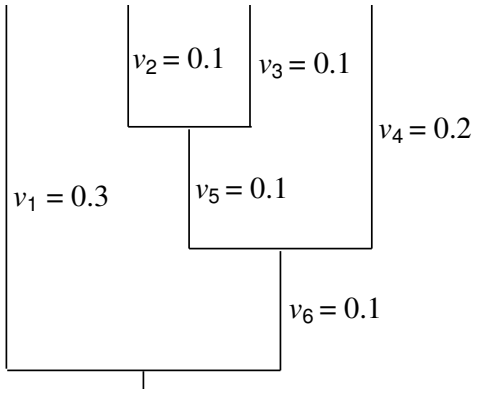

Figure 4: **The model tree for the simulations**. We simulate data on this tree. The branch lengths are denoted v_i.

In this section, we will test our knowledge obtained in the previous section by simulating DNA sequences on a phylogenetic tree. Simulation is often used in phylogenetics. Simulation has been used to elucidate the statistical properties of different phylogenetic methods, and it can be used to generate the null distribution of a test statistic in phylogenetic hypothesis testing. In other words, learning to simulate DNA sequences is not a wasted effort. Not only can you strengthen your intuition of phylogenetic methods, but you may also be able to apply simulation for your own research.

How exactly should one simulate evolution on a phylogeneitc tree? One basic point is that an alignment should be simulated on a site-by-site basis. That is, we first simulate the data at the first site, then the second site, and so on. We can take this approach because of the assumption of independence of the substitution process across sites; to simulate the data at a particular site (column in the alignment), we don't need to know the results of the simulation at any other site.

Another basic point is that we must know all of the parameters of the simulation: we have to decide on the precise phylogenetic tree on which to simulate the DNA sequences; we need to know the branch lengths on this tree; and, finally, we must pick a substitution model. The substitution model is a matrix of rates, specifying the rate of change from one nucleotide to another. In other words, we are taking a God-like view of the situation. We know *everything* about the evolutionary history and process. Of course, in reality we never know everything about how organisms evolved, but must make strong assumptions about how evolution occurred in order to estimate (make educated guesses) at the underlying evolutionary history. However, pretending to be a God, even for a little while, is a great feeling.

In the following, we will evolve DNA sequences on the four-taxon tree shown in Figure 4. We will also assume that DNA substitution occurs according to the HKY85 model with the parameters fixed to the following values: $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$. The rate matrix, then, is

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

Now, we are ready to simulate data on the tree of Figure 4. We will go over four different methods for simulating data, each of which takes advantage of our knowledge of continuous-time Markov chains.

2.1 Method 1

The first method only relies on our ability to generate exponentially distributed random numbers. If we generate a uniform random number on the interval (0,1), we can generate an exponential random number (with parameters λ) using the transformation $t = -\frac{1}{\lambda}\log_e(u)$ (where u is the uniform random number and t is the exponentially distributed random number). The first method involves an addition to the tree of Figure 4 which seems unusual: We take the tree of Figure 4, and add a 'tail' to it—a branch that extends

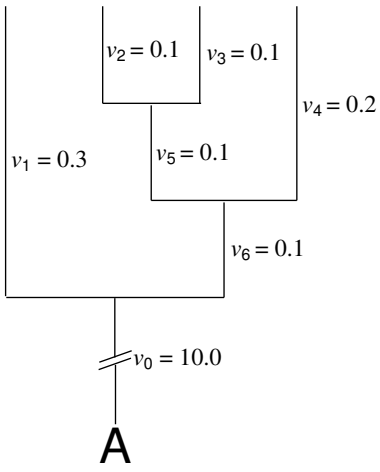

Figure 5: **The model tree for the simulations, with a tail**. The tree of Figure 4, with a long branch at the root that starts in nucleotide A.

for some distance from the root of the tree. In this case, the branch at the root of the tree is $v_0 = 10.0$ in length. Moreover, we assume that the process is in state (nucleotide) A at the very root of the tree. The situation we have is like that shown in Figure 5.

We simulate the process starting at the root of the tree. The process is in state A, meaning that the only relevant row of the rate matrix is the first one:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \end{pmatrix}$$

We wait an exponentially distributed amount of time with parameter $\lambda = 0.886$ until the next substitution occurs. When the substitution occurs, it is to a C, G, or T with probabilities $\frac{0.190}{0.886} = 0.214$, $\frac{0.633}{0.886} = 0.714$, and $\frac{0.063}{0.886} = 0.072$, respectively. In the first section, we used this method for simulating along a single branch of a tree. Here we apply the method with vigor, applying it to each branch in the tree from the root to the tips. We continue to simulate up the root branch of the tree until our simulation exceeds the length of the branch. We then record the nucleotide state the process was in when it exceeded a length of 10. We write this state at the end of the root branch, where it splits into branches 1 and 6. We then repeat the simulation process for branch 1 and then branch 6, recording the state the process is in at the end of those two branches. We then concentrate our attention on branches 4 and 5, and then on branches 2 and 3. At the end, we should have nucleotides at the ends of branches 1, 2, 3, and 4.

One puzzling aspect of this simulation is why we always start the process in nucleotide A, and why we even bothered to add the tail to the root of the tree. We did this because for Method 1, we only are going to allow ourselves to generate exponential random numbers. If this is the case, we can use our understanding of the rate matrix as specifying waiting times between substitutions to complete our simulation. However, we are not allowing ourselves knowledge of the stationary distribution of the substitution process. Hence, we always start our simulations in a particular nucleotide (in this case we chose to start in the nucleotide A), and then simulate the process for a long time along the root (tail) branch of the tree. The hope is that if we make the length of the tail branch long enough, that the process is at stationarity by the time it reaches the first split in the tree (the speciation event that eventually produces the four species at the tips of the tree).

Method 1 relies on the idea that we can come pretty near to stationarity with a moderately long branch. We know that the stationary distribution of the HKY85 process of nucleotide substitution with the specific parameters we chose is $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, $\pi_T = 0.1$. We also know that the transition probability

for a branch of $v = 10.0$ is

$$\mathbf{P}(10.0) = \begin{pmatrix} 0.401 & 0.299 & 0.200 & 0.099 \\ 0.399 & 0.301 & 0.199 & 0.100 \\ 0.401 & 0.299 & 0.200 & 0.099 \\ 0.399 & 0.301 & 0.199 & 0.100 \end{pmatrix}$$

The transition probability matrix tells us that if we start in nucleotide A, then we end up in state A, C, G, and T with probabilities 0.401, 0.299, 0.200, and 0.099, respectively. These numbers are very close to the actual stationary probabilities, so perhaps this method is not such a bad one.

2.2 Method 2

The second simulation method we explore is similar to the first one, but we eliminate the long tail on the tree. Instead of simulating the process along a long branch before we reach the first split in the tree, we simply decide which nucleotide is at the first split by sampling from the stationary distribution. We know that the stationary probabilities of the process are $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, $\pi_T = 0.1$, so why not simply pick a nucleotide at random with these probabilities?

2.3 Method 3

The third simulation method does away with the need to generate exponential random variables. It takes advantage of our knowledge of the stationary distribution (as does the second method), but also takes advantage of our ability to calculate transition probabilities. There are only three different lengths of branches on our model tree (0.1, 0.2, and 0.3). The transition probabilities are

$$\mathbf{P}(0.10) = \begin{pmatrix} 0.919 & 0.018 & 0.056 & 0.006 \\ 0.024 & 0.934 & 0.012 & 0.029 \\ 0.113 & 0.018 & 0.863 & 0.006 \\ 0.025 & 0.086 & 0.012 & 0.877 \end{pmatrix} \quad \mathbf{P}(0.20) = \begin{pmatrix} 0.851 & 0.035 & 0.100 & 0.011 \\ 0.047 & 0.876 & 0.023 & 0.052 \\ 0.201 & 0.035 & 0.750 & 0.011 \\ 0.047 & 0.156 & 0.023 & 0.771 \end{pmatrix}$$

$$\mathbf{P}(0.30) = \begin{pmatrix} 0.795 & 0.051 & 0.135 & 0.017 \\ 0.069 & 0.824 & 0.034 & 0.071 \\ 0.270 & 0.051 & 0.659 & 0.017 \\ 0.069 & 0.214 & 0.034 & 0.681 \end{pmatrix}$$

Instead of drawing exponential random variables, and generating the process continuously across the entire tree, our simulation jumps from node to node on the tree. First, we generate the nucleotide at the root of the tree (the first split leading to all four taxa) by drawing from the stationary distribution. Then, we use the transition probabilities to simulate from one end to the other of each branch on the tree. We start from the root of the tree, and simulate up the tree to progressively higher branches until we have simulated a nucleotide at each tip of the tree.

2.4 Method 4

The last method we will discuss involves calculating the probability of each possible pattern of nucleotides that we could observe at the tips of the tree. There are four tips, so there are $4^4 = 256$ possible patterns of nucleotides we could observe. We use the formula for the likelihood, discussed in the first section, to calculate the probability (likelihood) of each pattern.

We could quickly simulate the data at one site (column) in an alignment by generating one uniform random number on the interval (0,1). We would first decide which pattern would result from a particular random number by tabulating 256 intervals. For example, we might decide that if the uniform random number is between 0 and 0.199465, that the pattern AAAA is the result; that if the uniform random number is between 0.199465 and 0.20365 that the pattern AAAC results; that if the uniform random number is between 0.20365 and 0.218361 that the pattern AAAG results; and so on. These intervals are calculated directly from the numbers in Table 2. (If you do not see how this was done, here is a hint: $0.199465 + 0.004185 = 0.20365$ and $0.199465 + 0.004185 + 0.014711 = 0.218361$.)

This method can work well when the number of tips on the tree is small, keeping the number of possible patterns manageable. However, when the number of tips gets to be about 8 or 9, the number of possible patterns becomes too large to make this method a reasonable one for simulating data.

Table 2: **Probabilities of individual sites**. The probabilities of all 256 site patterns when simulated on the tree of Figure 4 under the HKY85 model of nucleotide substitution. The nucleotides for the site patterns are ordered from left to right (on the tree from Figure 4). The probabilities of all 256 site patterns are listed in Table 2. These probabilities were calculated on the tree of Figure 4 under the HKY85 model of substitution using our usual parameterization for the HKY85 model.

Pattern	Prob.	Pattern	Prob.	Pattern	Prob.	Pattern	Prob.
AAAA	0.199465	AGAA	0.014711	CAAA	0.018317	CGAA	0.001490
AAAC	0.004185	AGAC	0.000725	CAAC	0.000628	CGAC	0.000210
AAAG	0.014711	AGAG	0.019868	CAAG	0.001490	CGAG	0.002878
AAAT	0.001395	AGAT	0.000242	CAAT	0.000166	CGAT	0.000048
AACA	0.009075	AGCA	0.000843	CACA	0.005277	CGCA	0.000669
AACC	0.000703	AGCC	0.000315	CACC	0.004524	CGCC	0.002262
AACG	0.000843	AGCG	0.002202	CACG	0.000669	CGCG	0.002304
AACT	0.000121	AGCT	0.000048	CACT	0.000375	CGCT	0.000188
AAGA	0.028625	AGGA	0.005985	CAGA	0.003304	CGGA	0.001065
AAGC	0.000702	AGGC	0.000755	CAGC	0.000210	CGGC	0.000209
AAGG	0.005985	AGGG	0.032738	CAGG	0.001065	CGGG	0.006655
AAGT	0.000234	AGGT	0.000252	CAGT	0.000048	CGGT	0.000059
AATA	0.003025	AGTA	0.000281	CATA	0.000959	CGTA	0.000120
AATC	0.000121	AGTC	0.000048	CATC	0.000360	CGTC	0.000180
AATG	0.000281	AGTG	0.000734	CATG	0.000120	CGTG	0.000420
AATT	0.000154	AGTT	0.000073	CATT	0.000404	CGTT	0.000202
ACAA	0.004185	ATAA	0.001395	CCAA	0.000628	CTAA	0.000166
ACAC	0.005482	ATAC	0.000350	CCAC	0.009592	CTAC	0.000415
ACAG	0.000725	ATAG	0.000242	CCAG	0.000210	CTAG	0.000048
ACAT	0.000350	ATAT	0.001594	CCAT	0.000415	CTAT	0.001214
ACCA	0.000703	ATCA	0.000121	CCCA	0.004524	CTCA	0.000375
ACCC	0.019527	ATCC	0.000752	CCCC	0.167489	CTCC	0.005866
ACCG	0.000315	ATCG	0.000048	CCCG	0.002262	CTCG	0.000188
ACCT	0.000752	ATCT	0.001546	CCCT	0.005866	CTCT	0.007452
ACGA	0.000702	ATGA	0.000234	CCGA	0.000210	CTGA	0.000048
ACGC	0.001837	ATGC	0.000116	CCGC	0.004796	CTGC	0.000208
ACGG	0.000755	ATGG	0.000252	CCGG	0.000209	CTGG	0.000059
ACGT	0.000116	ATGT	0.000535	CCGT	0.000208	CTGT	0.000607
ACTA	0.000121	ATTA	0.000154	CCTA	0.000360	CTTA	0.000404
ACTC	0.001781	ATTC	0.000517	CCTC	0.011625	CTTC	0.001716
ACTG	0.000048	ATTG	0.000073	CCTG	0.000180	CTTG	0.000202
ACTT	0.000517	ATTT	0.004711	CCTT	0.001716	CTTT	0.013873

Table 2: **Probabilities of individual sites, continued**.

Pattern	Prob.	Pattern	Prob.	Pattern	Prob.	Pattern	Prob.
GAAA	0.045565	GGAA	0.005060	TAAA	0.006106	TGAA	0.000497
GAAC	0.001004	GGAC	0.000453	TAAC	0.000166	TGAC	0.000048
GAAG	0.005060	GGAG	0.017648	TAAG	0.000497	TGAG	0.000959
GAAT	0.000335	GGAT	0.000151	TAAT	0.000099	TGAT	0.000038
GACA	0.002514	GGCA	0.000532	TACA	0.000959	TGCA	0.000120
GACC	0.000315	GGCC	0.000194	TACC	0.000548	TGCC	0.000274
GACG	0.000532	GGCG	0.002904	TACG	0.000120	TGCG	0.000420
GACT	0.000048	GGCT	0.000036	TACT	0.000215	TGCT	0.000108
GAGA	0.014437	GGGA	0.008240	TAGA	0.001101	TGGA	0.000355
GAGC	0.000476	GGGC	0.001251	TAGC	0.000048	TGGC	0.000059
GAGG	0.008240	GGGG	0.056794	TAGG	0.000355	TGGG	0.002218
GAGT	0.000159	GGGT	0.000417	TAGT	0.000038	TGGT	0.000030
GATA	0.000838	GGTA	0.000177	TATA	0.001119	TGTA	0.000143
GATC	0.000048	GGTC	0.000036	TATC	0.000231	TGTC	0.000116
GATG	0.000177	GGTG	0.000968	TATG	0.000143	TGTG	0.000488
GATT	0.000073	GGTT	0.000040	TATT	0.000893	TGTT	0.000447
GCAA	0.001004	GTAA	0.000335	TCAA	0.000166	TTAA	0.000099
GCAC	0.001837	GTAC	0.000116	TCAC	0.001389	TTAC	0.000240
GCAG	0.000453	GTAG	0.000151	TCAG	0.000048	TTAG	0.000038
GCAT	0.000116	GTAT	0.000535	TCAT	0.000240	TTAT	0.002009
GCCA	0.000315	GTCA	0.000048	TCCA	0.000548	TTCA	0.000215
GCCC	0.009764	GTCC	0.000376	TCCC	0.019456	TTCC	0.001275
GCCG	0.000194	GTCG	0.000036	TCCG	0.000274	TTCG	0.000108
GCCT	0.000376	GTCT	0.000773	TCCT	0.001275	TTCT	0.006924
GCGA	0.000476	GTGA	0.000159	TCGA	0.000048	TTGA	0.000038
GCGC	0.001823	GTGC	0.000117	TCGC	0.000694	TTGC	0.000120
GCGG	0.001251	GTGG	0.000417	TCGG	0.000059	TTGG	0.000030
GCGT	0.000117	GTGT	0.000530	TCGT	0.000120	TTGT	0.001005
GCTA	0.000048	GTTA	0.000073	TCTA	0.000231	TTTA	0.000893
GCTC	0.000891	GTTC	0.000258	TCTC	0.004935	TTTC	0.003240
GCTG	0.000036	GTTG	0.000040	TCTG	0.000116	TTTG	0.000447
GCTT	0.000258	GTTT	0.002355	TCTT	0.003240	TTTT	0.031522