# The Idiot's Guide to the Zen of Likelihood in a Nutshell in Seven Days for Dummies, Unleashed

## A gentle introduction, for those of us who are small of brain, to the calculation of the likelihood of molecular sequences

Peter G. Foster[*]

July 28, 2001

This informal explanation is made for phylogeneticists who want to know what is inside the black box, who want to know where the numbers come from. Perhaps you have given yourselves a similar explanation for how parsimony reconstruction works. You might have made up some data for a few fictitious taxa, and tried to fit them to a couple of different trees, and satisfied yourself that one tree is more parsimonious than another. Perhaps you then actually typed in the data and fed them to paup or protpars to check the tree lengths that you calculated against those calculated by the program. Knowing that you and the program calculate the same way is satisfying, and makes you feel better about using the program for more sizable data. It is to this sort of person that this explanation is directed. Although simple demo parsimony calculations can be done by hand with pen and paper, simple demo likelihood calculations are more involved, and the rationale more elusive.

The likelihood is probability of the data given the model. Why not then just call it the probability? I suppose it was given the special new name to emphasize that you are talking about the probability of data, not the probability of some abstract event happening, and also to emphasize that it is based on a model.

The data have already been collected. They don't change—they are a given. What you can change, however, is the model. The model is the picture of the way that you think that things work. It is an idea, and so you can do whatever you want with it.

Lets say that you flip a coin and get a head. That's the datum. Now if you think that its a fair coin (the model), the datum will have a probability of 1/2. But if your model is that you have a two-headed coin, then your datum will have a probability of 1. The lesson here is that the model you have in mind can have a big effect on the probability of the data.

For molecular evolution, the data are an alignment of sequences, and the model in its large sense is the tree that relates the sequences plus the mechanism of molecular change. The tree and the mechanism together are your idea of the way that you think that things work. But it is usual to separate the two parts of the model, and call the tree part "the tree" and the mechanism part "the model". So the definition of the model is somewhat loose. We will keep it that way.

The mechanism part of the model—lets call it the model, in keeping with our loose definition—is an idea of the way you think that molecular sequences change over time. Whereas the driving force behind morphological change is selection, it appears that molecules change, for the most part, in a random way. By saying that we

[*] Department of Zoology, The Natural History Museum. email: p.foster@nhm.ac.uk

think it is random doesn't mean that we think that everything proceeds equally—we may be dealing with loaded dice.

Lets just speak of DNA models. I imagine the model as having two parts: the composition and the process. The composition is just the proportion of the four nucleotides. For example you might think that the four bases are in equal proportions, or that there are twice as many $a$'s as $c$'s. Or you might let the data decide for you, and so if your sequences are $g$ and $c$-rich, that is what the composition part of the model gets.

*The likelihood of a sequence*

Lets do our first likelihood calculations. We are going to calculate the probability of the nucleotide "$a$". Just one sequence, one nucleotide long, with no tree involved. Since there is no need for nucleotide change, we don't need the process part of the model, we just need the composition part. If our model is that everything is 100% $a$, then the likelihood of $a$ is 1. If our model is that everything is 100% $c$, then the likelihood of $a$ is zero. (This might be a tip-off that the model does not fit the data well). If our model is that $a$ has a composition of 33%, then the likelihood of $a$ is 0.33.

Lets calculate the likelihood of a single sequence with two nucleotides, say "$ac$". If the model is the Jukes-Cantor model, which has a composition of 1/4 for each base, then the likelihood will be $1/4 \times 1/4 = 1/16$. If the model has a composition of 40% $a$ and 10% $c$, the likelihood of the sequence will be $0.4 \times 0.1 = 0.04$.

If you take the 16 possible dinucleotides and calculate the likelihood for all of them, the sum of those likelihoods should be 1. If you choose to evaluate their likelihoods with the JC model, its $1/16 \times 16$, but it should be true of whatever model you choose. This should always be true: the sum of the likelihoods of all the different data possibilities should be 1.

*The likelihood of a one-branch tree*

The other part of the model, the process part, is needed if we have more than one sequence related by a tree. The process might be described by sentences, or by equations, or by a matrix of numbers, describing how the nucleotides change from one to another. Lets use a very simple alignment with only two sequences, each one base long, an $a$ in one sequence and a $c$ in the other. The sequences are related by a simple tree, in this case a single branch. Lets evaluate the likelihood of this tree under a model that says that the composition is 1/4 $a$ and 1/4 $c$, and the process part of my model is "the probability of an $a$ changing to a $c$, or vice versa, is 0.4". So, starting with $a$, the probability of the sequence $a$ is 1/4, and the probability of the branch is 0.4. The probability of the two together, the tree, is $1/4 \times 0.4 = 0.1$. If we start with the $c$ we get the same, because the model is reversible.

There are 16 possible changes from one nucleotide to another (including remaining itself), and we can organize the probabilities of those changes in a $4 \times 4$ matrix, eg

$$\boldsymbol{P} = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$

I'll always use the convention that the order of the bases is $a$, $c$, $g$, and $t$, alphabetical order. So this says that the probability of an $a$ changing to a $c$ is 0.01, and the probability of a $c$ remaining a $c$ is 0.983, $\boldsymbol{P}_{t \to g} = 0.005$ and so on. Here the rows sum to 1, which says that the probability of something happening is 1, a comforting thought. The columns don't sum to anything in particular, however. The composition part of the model I'll denote with a $\pi$, as in $\pi = [0.1, 0.4, 0.2, 0.3]$, with the same alphabetical ordering of the bases.

Armed with the model stated this way, you can calculate the likelihood of a one branch tree between two sequences. Lets say that we have an alignment

$$\begin{array}{cccc} c & c & a & t \\ c & c & g & t \end{array}$$

the likelihood, going from the first to the second sequence, will be

$$\begin{aligned} &= \pi_c \; \boldsymbol{P}_{c\to c} \; \pi_c \; \boldsymbol{P}_{c\to c} \; \pi_a \; \boldsymbol{P}_{a\to g} \; \pi_t \; \boldsymbol{P}_{t\to t} \\ &= 0.4 \times 0.983 \times 0.4 \times 0.983 \\ &\qquad \times 0.1 \times 0.007 \times 0.3 \times 0.979 \\ &= 0.0000300 \end{aligned}$$

*Different branch lengths*

Of course the model above doesn't take into account the possibility of different branch lengths. We would think that for very short branch lengths, the probability of a base changing to another is low, and the probability of it staying the same is high, near one. The matrix $\boldsymbol{P}$ above seems to describe a short branch. For long branches, the probability of a base staying the same would drop, and the probability that it changes to something else would rise.

Lets say that the matrix above describes a branch with a length of a Certain Evolutionary Distance, or ced unit. So the likelihood that we calculated was for 1 ced. What would the likelihood be for the same alignment with a branch of 2 ced? We can get the probabilities for that by multiplying the matrix by itself.

$$\begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}^2 = \begin{bmatrix} 0.953 & 0.02 & 0.013 & 0.015 \\ 0.005 & 0.966 & 0.01 & 0.02 \\ 0.007 & 0.02 & 0.959 & 0.015 \\ 0.005 & 0.026 & 0.01 & 0.959 \end{bmatrix}$$

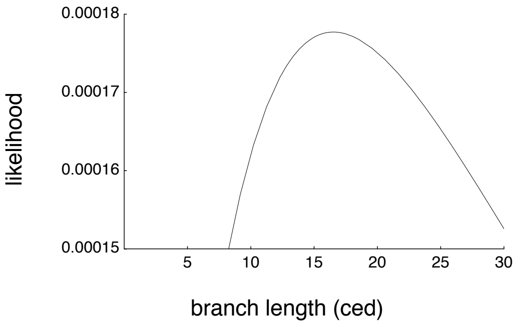
which gives a likelihood of 0.0000559. Similarly for 3 ced the probability matrix is

$$\boldsymbol{P}^3 = \begin{bmatrix} 0.93 & 0.029 & 0.019 & 0.022 \\ 0.007 & 0.949 & 0.015 & 0.029 \\ 0.01 & 0.029 & 0.939 & 0.022 \\ 0.007 & 0.038 & 0.015 & 0.94 \end{bmatrix}$$

which gives a likelihood of 0.0000782.

Notice that as the branch length increases the probabilities on the diagonal are going down and the probabilities off diagonal are going up. Notice also that the likelihood of the tree is going up as we go from 1 to 3 ced units. What happens if we keep going?

| branch length<br>(ced units) | likelihood |
|---|---|
| 1 | 0.0000300 |
| 2 | 0.0000559 |
| 3 | 0.0000782 |
| 10 | 0.000162 |
| 15 | 0.000177 |
| 20 | 0.000175 |
| 30 | 0.000152 |



The likelihood rises to a maximum somewhere between 10 and 20 ced units.

If you raise $\boldsymbol{P}$ to a very large power, $\pi$ pops out. So $\pi$ was actually built-in to the probability matrix $\boldsymbol{P}$.

$$\boldsymbol{P}^{10^6} = \begin{bmatrix} 0.1 & 0.4 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.2 & 0.3 \end{bmatrix}$$

*Rate matrices*

If you want to calculate $5^4$, you can get that by $\exp(4\log(5))$. In a similar way you can power matrices by taking the matrix log to get the rate matrix, multiplying the rate matrix by the branch length, and then taking the matrix exponent of the product. This way you can get non-integral exponents, and there are other advantages as well. If you go this route, you can fully separate the composition from the process (we saw above that the composition was built-in to the probability matrices). Another good reason to do this is so that you can easily express your branch lengths in substitutions per site, rather than arbitrary units such as ced units used above, and furthermore you can get branch lengths all the way from zero to infinity.

If we use our example $\boldsymbol{P}$ from above,

$$\log \boldsymbol{P} = \begin{bmatrix} -0.0244 & 0.0101 & 0.0067 & 0.0076 \\ 0.0025 & -0.0176 & 0.005 & 0.0101 \\ 0.0034 & 0.0101 & -0.021 & 0.0076 \\ 0.0025 & 0.0134 & 0.005 & -0.021 \end{bmatrix}$$

Here the sum of each row is zero. This matrix corresponds to one ced, and if we take the exponential we recover the matrix corresponding to one ced. What we want though, is a matrix such that if we take its exponential we get a $\boldsymbol{P}$ corresponding to one substitution per site. That is done by scaling $\log \boldsymbol{P}$ so that when the rows of $\log \boldsymbol{P}$ are multiplied by $\pi_{row}$ the off-diagonal elements sum to 1. The resulting scaled $\log \boldsymbol{P}$, which I will call $\boldsymbol{Q}$, when its exponent is taken gives a $\boldsymbol{P}$ corresponding to 1 substitution per site. In general,

$$e^{\boldsymbol{Q}\nu} = \boldsymbol{P}(\nu)$$

for branch length $\nu$.

If we scale the $\log \boldsymbol{P}$ above appropriately, by a factor of 50, we get

$$\boldsymbol{Q} = \begin{bmatrix} -1.218 & 0.504 & 0.336 & 0.378 \\ 0.126 & -0.882 & 0.252 & 0.504 \\ 0.168 & 0.504 & -1.05 & 0.378 \\ 0.126 & 0.672 & 0.252 & -1.05 \end{bmatrix}$$

Now if we multiply big $\Pi$, a diagonal matrix of the $\pi$ elements, by $\boldsymbol{Q}$

$$\begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.3 \end{bmatrix} \cdot \boldsymbol{Q} =$$

$$\begin{bmatrix} -0.122 & 0.05 & 0.034 & 0.038 \\ 0.05 & -0.353 & 0.101 & 0.202 \\ 0.034 & 0.101 & -0.21 & 0.076 \\ 0.038 & 0.202 & 0.076 & -0.315 \end{bmatrix}$$

we get a matrix where the off-diagonal elements sum to 1, and the diagonal elements sum to -1. When this is the case, $\boldsymbol{P}$'s generated from this $\boldsymbol{Q}$ will give branch lengths in substitutions per site.

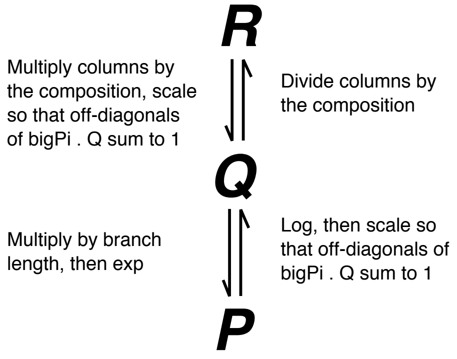
*Separating the composition from the rates*

If we divide the columns of $\boldsymbol{Q}$ by $\pi_{col}$, the composition is separated from the rates. Then you can for example use exactly the same rate matrix with different compositions. The rate matrix $\boldsymbol{R}$ for the model that we have been using is

$$\boldsymbol{R} = \begin{bmatrix} - & 0.3 & 0.4 & 0.3 \\ 0.3 & - & 0.3 & 0.4 \\ 0.4 & 0.3 & - & 0.3 \\ 0.3 & 0.4 & 0.3 & - \end{bmatrix}$$

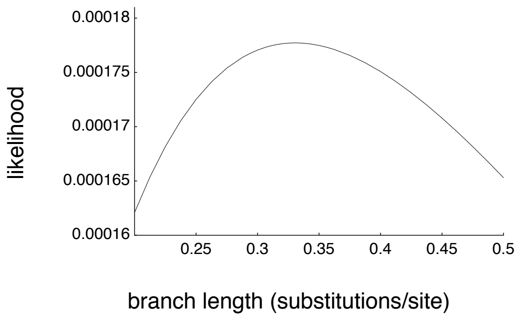
The diagonal elements don't matter in this context, so they are left out. The scale doesn't matter either. However, for a reversible model it should be symmetrical. When PAUP expresses a rate matrix with the lset rmatrix subcommand, it uses 5 numbers ($a \rightarrow c$, $a \rightarrow g$, $a \rightarrow t$, $c \rightarrow g$, $c \rightarrow t$), scaling so that the sixth is 1.0. PAUP would express that $\boldsymbol{R}$ by rmatrix=(1.0 1.33333 1.0 1.0 1.333333)

Interconversions between $\boldsymbol{R}$, $\boldsymbol{Q}$ and $\boldsymbol{P}$ are summarized here. In practice, you would hardly ever go from $\boldsymbol{P}$ to $\boldsymbol{Q}$, or from $\boldsymbol{Q}$ to $\boldsymbol{R}$.


R
Multiply columns by the composition, scale so that off-diagonals of bigPi . Q sum to 1
Divide columns by the composition
Q
Multiply by branch length, then exp
Log, then scale so that off-diagonals of bigPi . Q sum to 1
P

*The maximum likelihood branch length in substitutions per site*

The likelihood of the alignment *ccat* and *ccgt* at various distances is



The maximum can be found numerically, by successive approximation. It is at a branch length of 0.330614, at a likelihood of 0.0001777.

*Checking that PAUP* gets the right answer*

```
#NEXUS

begin data;
    dimensions ntax=2 nchar=4;
    format datatype=dna;
    matrix
    A   ccat
    B   ccgt
    ;
end;

begin paup;
    set criterion=distance;
    lset
        nst=6
        rmatrix = (1.0 1.33333 1.0
                    1.0 1.333333)
        basefreq = (0.1 0.4 0.2)
        ;
    dset distance=ml;
    showdist;  [got 0.33061]
end;
```

*A two-branch tree*

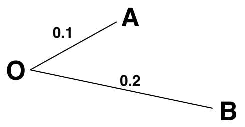
First, for the $\boldsymbol{Q}$ above, the corresponding $\boldsymbol{P}$ matrices for 0.1, 0.2, and 0.3 substitutions per site are

$$\boldsymbol{P}(0.1) = \begin{bmatrix} 0.886 & 0.047 & 0.031 & 0.036 \\ 0.012 & 0.918 & 0.024 & 0.046 \\ 0.016 & 0.047 & 0.902 & 0.036 \\ 0.012 & 0.062 & 0.024 & 0.902 \end{bmatrix}$$

$$\boldsymbol{P}(0.2) = \begin{bmatrix} 0.787 & 0.089 & 0.057 & 0.067 \\ 0.022 & 0.847 & 0.045 & 0.086 \\ 0.029 & 0.089 & 0.815 & 0.067 \\ 0.022 & 0.115 & 0.045 & 0.819 \end{bmatrix}$$

$$\boldsymbol{P}(0.3) = \begin{bmatrix} 0.7 & 0.126 & 0.08 & 0.094 \\ 0.031 & 0.786 & 0.063 & 0.119 \\ 0.04 & 0.126 & 0.74 & 0.094 \\ 0.031 & 0.159 & 0.063 & 0.747 \end{bmatrix}$$

We have calculated the likelihood of a one branch tree; lets move on to two. We will use the same model, using $\boldsymbol{P}(0.1)$, $\boldsymbol{P}(0.2)$ and $\boldsymbol{P}(0.3)$ that we just made, and the alignment from before. The tree has taxon A with sequence *ccat* and taxon B with sequence *ccgt* arranged as



where O is the origin, the root, and the numbers are the branch lengths. We will calculate the likelihood three ways:

*Way 1: From A to B in one step*

This is just like a one-branch calculation. For this we use the $\boldsymbol{P}(0.3)$ matrix, because it is a distance of 0.3 from A to B. The likelihood is

$$\begin{aligned} &= \pi_c \, \boldsymbol{P}_{c\to c} \, \pi_c \, \boldsymbol{P}_{c\to c} \, \pi_a \, \boldsymbol{P}_{a\to g} \, \pi_t \, \boldsymbol{P}_{t\to t} \\ &= 0.4 \times 0.786 \times 0.4 \times 0.786 \\ &\quad \times 0.1 \times 0.08 \times 0.3 \times 0.747 \\ &= 0.000177 \end{aligned}$$

*Way 2: From A to B in two steps*

For the first part, from A to O, we use $\boldsymbol{P}(0.1)$. We include the $\pi$ values in this part, because we are starting with A. For the first site in the A sequence, $c$, it will be $\pi_c\times$ ...what? We don't know what the O sequence is at that site. It is most probably a $c$, but it could be anything. The probability for the branch from A to O is the sum of the 4 possibilities.

$$
\begin{aligned}
&= \pi_c\ \boldsymbol{P}_{c\to a} + \pi_c\ \boldsymbol{P}_{c\to c} + \pi_c\ \boldsymbol{P}_{c\to g} + \pi_c\ \boldsymbol{P}_{c\to t} \\
&= 0.4 \times 0.012 + 0.4 \times 0.918 \\
&\qquad + 0.4 \times 0.024 + 0.4 \times 0.046 \\
&= 0.4 \\
&= \pi_c
\end{aligned}
$$

When we add the second branch, from O to B, into the calculation, we don't need to put in more $\pi$ terms—we only need those once, at the starting place. When we are summing over the possibilities of what the unknown sequence at O can be, whatever the unknown is at O for the branch from A to O, it will be the same unknown at O from O to B. So we are still summing over only four possibilities. The calculation for the first site will be as above, but now adding in the branch from O to B using $\boldsymbol{P}(0.2)$, we have the likelihood as the sum over the four possibilities

$$
\begin{aligned}
&= \pi_c\ \boldsymbol{P}_{0.1,c\to a}\ \boldsymbol{P}_{0.2,a\to c} + \pi_c\ \boldsymbol{P}_{0.1,c\to c}\ \boldsymbol{P}_{0.2,c\to c} + \\
&\qquad \pi_c\ \boldsymbol{P}_{0.1,c\to g}\ \boldsymbol{P}_{0.2,g\to c} + \pi_c\ \boldsymbol{P}_{0.1,c\to t}\ \boldsymbol{P}_{0.2,t\to c} \\
&= 0.4 \times 0.012 \times 0.089 + 0.4 \times 0.918 \times 0.847 \\
&\qquad + 0.4 \times 0.024 \times 0.089 + 0.4 \times 0.046 \times 0.115 \\
&= 0.3145 \\
&= \pi_c\ \boldsymbol{P}_{0.3,c\to c} \\
&= 0.4 \times 0.786
\end{aligned}
$$

The likelihood for the other sites are calculated in a similar way, and the product of the site likelihoods is 0.000177.

*Way 3: In two parts, starting from O*

When we start with O, since we don't know what it is, we add up the probabilities of the four possibilities. Since we are starting with O, the $\pi$'s refer to that node. The likelihood for the first position, $c$ with $c$, is

$$
\begin{aligned}
&= \pi_a\ \boldsymbol{P}_{0.1,a\to c}\ \boldsymbol{P}_{0.2,a\to c} + \pi_c\ \boldsymbol{P}_{0.1,c\to c}\ \boldsymbol{P}_{0.2,c\to c} + \\
&\qquad \pi_g\ \boldsymbol{P}_{0.1,g\to c}\ \boldsymbol{P}_{0.2,g\to c} + \pi_t\ \boldsymbol{P}_{0.1,t\to c}\ \boldsymbol{P}_{0.2,t\to c} \\
&= 0.1 \times 0.047 \times 0.089 + 0.4 \times 0.918 \times 0.847 \\
&\qquad + 0.2 \times 0.047 \times 0.089 + 0.3 \times 0.062 \times 0.115 \\
&= 0.3145
\end{aligned}
$$

What we are asking for is the likelihood of the data, *cc*. We can easily calculate probabilities of the four possible ways that that data might have come into being. To get the likelihood of the data at that site we sum the probabilities of those possibilities. The product of the 4 site likelihoods is 0.000177.

The two-part calculations look like a complicated way to calculate something that could have been done simply by straightening out the two branches into one long one. Granted, for a two branch tree this is true. However, the next example uses a three branch tree, and it could not be done using the branch-straightening method—it needs the complicated way.

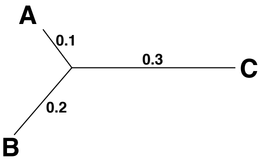
One lesson for this part is that it doesn't matter where you start, ie where you put the root: you will still get the same answer. We put the root at A and at O. You could put the root at B or half-way between A and O, or anywhere, and it will still work. This is Felsenstein's "Pulley Principle".

*A three branch tree*

Lets use an alignment

|   |   |
|---|---|
| A | `CCAT` |
| B | `CCGT` |
| C | `GCAT` |

arranged as



We root the tree at the internal node, and start the likelihood calculations there, as in Way 3, above. The likelihood of the first site is

$$\begin{aligned}
&= \pi_a \, \boldsymbol{P}_{0.1,a\to c} \, \boldsymbol{P}_{0.2,a\to c} \, \boldsymbol{P}_{0.3,a\to g} + \\
&\quad \pi_c \, \boldsymbol{P}_{0.1,c\to c} \, \boldsymbol{P}_{0.2,c\to c} \, \boldsymbol{P}_{0.3,c\to g} + \\
&\quad \pi_g \, \boldsymbol{P}_{0.1,g\to c} \, \boldsymbol{P}_{0.2,g\to c} \, \boldsymbol{P}_{0.3,g\to g} + \\
&\quad \pi_t \, \boldsymbol{P}_{0.1,t\to c} \, \boldsymbol{P}_{0.2,t\to c} \, \boldsymbol{P}_{0.3,t\to g} \\
&= 0.1 \times 0.047 \times 0.089 \times 0.08 + \\
&\quad 0.4 \times 0.918 \times 0.847 \times 0.063 + \\
&\quad 0.2 \times 0.047 \times 0.089 \times 0.74 + \\
&\quad 0.3 \times 0.062 \times 0.115 \times 0.063 \\
&= 0.0204
\end{aligned}$$

If we calculate the other three sites in a similar way, we get site likelihoods 0.245, 0.00368, and 0.166. If we multiply them together, we get a likelihood for the tree of $3.04 \times 10^{-6}$.

*Check with PAUP**

```
#NEXUS

begin paup;
    set storebrlens=yes;
end;

begin data;
    dimensions ntax=3 nchar=4;
    format datatype=dna;
    matrix
    A   ccat
    B   ccgt
    C   gcat
    ;
end;

begin trees;
    tree t1 = [&U] (A:0.1, B:0.2, C:0.3);
end;

begin paup;
    set criterion=likelihood;
    lset
        userbrlens=yes
        nst=6
        rmatrix = (1.0 1.33333 1.0
                    1.0 1.333333)
        basefreq = (0.1 0.4 0.2)
        ;
    lscores /sitelikes=yes;

end;
    [got site likes
     0.020385    0.245121
     0.003675    0.165724
     -log like = 12.70253,
     exp of which is 3.0434e-06 ]
```

*Selection, and slow and fast sites*

Noting that molecules change, for the most part, in a random way is a good place to start. However, a purely neutral model has a problem: if there is any natural selection it does not adequately reflect reality. Only occasionally, such as for example in pseudogenes, is the evolution of sequences adequately described by a purely neutral model. There is usually some selection of some parts of molecular sequences. Some sites are invariant, and the remaining sites vary to varying degrees. A trivial example of an invariant site is the start codon—it is under heavy selection and so neutral evolution does not apply.

We can complicate our models to allow for site rate heterogeneity, which is a way to deal with selection. For observed constant sites, we can say that there is a certain probability that the data may have arisen by that site being an invariable site, fixed in place by selection, or the data may have arisen because it is a possibly variable site that by chance has not varied yet. Although we could calculate the probability of either, we can't know which, and so we sum the probabilities of the possibilities. This summing is compounded on the summing that is done over the 4 possible bases at internal nodes. We can complicate matters further by saying in our model that if it is a variable site, then it might be evolving at a fast or slow rate, perhaps described by a discreet gamma distribution. Again, the strategy is to sum over the possibilities. Implementation details are left as an exercise for the reader.

# Appendix

Calculations here were done with *Mathematica*. It has MatrixPower and MatrixExp functions, but no MatrixLog function. An adequate stand-in is

```
    matrixLog[mat_] := Module[{dim},
        dim = Dimensions[mat][[1]];
        Sum[MatrixPower[mat -
            IdentityMatrix[dim], i]/
            (((-1)^(i + 1)) i), {i, 1, 50}]]
```

which uses a Taylor series. A working program would use eigensystems to exponentiate or calculate the logarithm of rate matrices.

The function for finding the maximum was the Golden Ratio method coded into *Mathematica.*