PhyloMath Lecture 8

by Dan Vanderpool 22 March, 2004

Topics of Discussion

- Transition:Transversion rate ratio (Kappa) vs. Transition:Transversion ratio (T-ratio)
- Calculating the expected number of substitutions using matrix algebra
- Why the General Time Reversible model can only have 5 relative rates
- Likelihood of a 5 taxon tree (next class will perform this calculation but take account of rate heterogeneity among sites)

The transition:transversion ***rate*** ratio vs. the T-ratio.

It is often observed in real datasets that transitions (Ti) occur at a rate different from (and often faster than) transversions (Tv), resulting in a Ti/Tv bias. There are multiple ways for nucleotide substitution models to account for this bias.

Some nucleotide substitution models account for Ti/Tv bias by calculating the Ti:Tv **rate** ratio. This ratio is defined as the rate at which transitions are occurring divided by the rate at which transversions are occurring. This ratio is often expressed as the Greek letter κ (kappa) as it is implemented in the K80 (K2P, Kimura Two Parameter)[1]nucleotide substitution model.

The Q-matrix for the K80 model looks like this

$$Q=\begin{array}{cccc} A & C & G & T \\ \left[\begin{array}{cccc} -\beta(\kappa+2) & \beta & \kappa\beta & \beta \\ \beta & -\beta(\kappa+2) & \beta & \kappa\beta \\ \kappa\beta & \beta & -\beta(\kappa+2) & \beta \\ \beta & \kappa\beta & \beta & -\beta(\kappa+2) \end{array}\right] \end{array}$$

Observe that the instantaneous rate at which all tranversions are occurring is equal to β.

Tv rate = β

[1] Maxi has referenced the publications in which these models were first described in the notes for lecture 7.

The instantaneous rate at which all transitions are occurring is β, modified by some number κ

Ti rate = κβ

To determine rate ratio we divide Ti rate by Ts rate and the β terms cancel

$$\frac{\text{Ti rate}}{\text{Ts rate}} = \frac{\kappa\cancel{\beta}}{\cancel{\beta}} = \kappa$$ leaving κ

This is one way to account for the transition/transversion bias observed in a data set.

The T-ratio is defined as the probability of any transition occurring in a single instant of time, *dt* (an interval during which only 0 or 1 substitutions can occur) divided by the probability of any transversion occurring in a single instant of time *dt*.

$$\frac{\text{Pr(Any Ti occurring in instant of time dt)}}{\text{Pr(Any Tv occurring in instant of time dt)}} =$$

$$\frac{\tfrac{1}{4}\mathrm{Pr}_{AG}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{CT}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{GA}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{TC}(\mathrm{dt})}{\tfrac{1}{4}\mathrm{Pr}_{AC}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{AT}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{CA}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{CG}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{GC}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{GT}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{TA}(\mathrm{dt})+\tfrac{1}{4}\mathrm{Pr}_{TG}(\mathrm{dt})} =$$

If we solve for this equation in terms of the Ti/Tv RATE! ratio from the previous Q - matrix this =

$$\frac{\tfrac{1}{4}\kappa\beta\,\mathrm{dt}+\tfrac{1}{4}\kappa\beta\,\mathrm{dt}+\tfrac{1}{4}\kappa\beta\,\mathrm{dt}+\tfrac{1}{4}\kappa\beta\,\mathrm{dt}}{\tfrac{1}{4}\beta\,\mathrm{dt}+\tfrac{1}{4}\beta\,\mathrm{dt}+\tfrac{1}{4}\beta\,\mathrm{dt}+\tfrac{1}{4}\beta\,\mathrm{dt}+\tfrac{1}{4}\beta\,\mathrm{dt}+\tfrac{1}{4}\beta\,\mathrm{dt}+\tfrac{1}{4}\beta\,\mathrm{dt}+\tfrac{1}{4}\beta\,\mathrm{dt}} =$$

$$\frac{4\kappa\beta\mathrm{dt}}{8\beta\mathrm{dt}} = \frac{\kappa\beta\mathrm{dt}}{2\beta\mathrm{dt}} \Rightarrow \text{terms cancel} \Rightarrow \frac{\kappa}{2}$$

Where the 1/4 term above results from the assumption of equal base frequencies in the K80 model.

What this demonstrates is that the quantity known as the transition:transversion ratio (tratio in PAUP*) is only 1/2 the value of the kappa parameter defined as part of the K80 model. So even though the "transition:transversion ratio" and the "transition:transversion rate ratio" sound like the same quantity, they measure different aspects of transition/transversion bias and can have quite different numerical values.

By looking at the rate matrix for the HKY85 model we see that κ must be set equal to 1 (one) in order for the model to be equivalent to the F81 model (in which rate of substitution is determined entirely by the nucleotide composition).

HKY85 rate matrix:

$$Q=\begin{bmatrix} -\beta(\pi_{\mathrm{Y}}+\kappa\beta\pi_{\mathrm{G}}) & \pi_{\mathrm{C}}\beta & \pi_{\mathrm{G}}\kappa\beta & \pi_{\mathrm{T}}\beta \\ \pi_{\mathrm{A}}\beta & -\beta(\pi_{\mathrm{R}}+\kappa\beta\pi_{\mathrm{T}}) & \pi_{\mathrm{G}}\beta & \pi_{\mathrm{T}}\kappa\beta \\ \pi_{\mathrm{A}}\kappa\beta & \pi_{\mathrm{C}}\beta & -\beta(\pi_{\mathrm{Y}}+\kappa\beta\pi_{\mathrm{A}}) & \pi_{\mathrm{T}}\beta \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\kappa\beta & \pi_{\mathrm{G}}\beta & -\beta(\pi_{\mathrm{R}}+\kappa\beta\pi_{\mathrm{C}}) \end{bmatrix}$$

(From the HKY85 rate matrix observe that if κ=1, (1)β= β meaning rates are equal)

For the F84 model to be equal to the F81 model, the kappa term used in the rate matrix must equal 0 (zero).

F84 rate matrix:

$$Q=\begin{bmatrix} -\beta(1-\pi_A)-\kappa\beta\left(\frac{\pi_{\mathrm{G}}}{\pi_{\mathrm{R}}}\right) & \pi_{\mathrm{C}}\beta & \pi_{\mathrm{G}}\beta+\left(\frac{\pi_{\mathrm{G}}}{\pi_{\mathrm{A}}+\pi_{\mathrm{G}}}\right)\kappa\beta & \pi_{\mathrm{T}}\beta \\ \pi_{\mathrm{A}}\beta & -\beta(1-\pi_C)-\kappa\beta\left(\frac{\pi_{\mathrm{T}}}{\pi_{\mathrm{Y}}}\right) & \pi_{\mathrm{G}}\beta & \pi_{\mathrm{T}}\beta+\left(\frac{\pi_{\mathrm{T}}}{\pi_{\mathrm{C}}+\pi_{\mathrm{T}}}\right)\kappa\beta \\ \pi_{\mathrm{A}}\beta+\left(\frac{\pi_{\mathrm{A}}}{\pi_{\mathrm{A}}+\pi_{\mathrm{G}}}\right)\kappa\beta & \pi_{\mathrm{C}}\beta & -\beta(1-\pi_G)-\kappa\beta\left(\frac{\pi_{\mathrm{A}}}{\pi_{\mathrm{R}}}\right) & \pi_{\mathrm{T}}\beta \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\beta+\left(\frac{\pi_{\mathrm{C}}}{\pi_{\mathrm{C}}+\pi_{\mathrm{T}}}\right)\kappa\beta & \pi_{\mathrm{G}}\beta & -\beta(1-\pi_{\mathrm{T}})-\kappa\beta\left(\frac{\pi_{\mathrm{C}}}{\pi_{\mathrm{Y}}}\right) \end{bmatrix}$$

Refer to the F84 rate matrix and notice that, for example, when κ=0 in the $A \to G$ transition, $\pi_{\mathrm{G}}\beta+\left(\frac{\pi_{\mathrm{G}}}{\pi_{\mathrm{A}}+\pi_{\mathrm{G}}}\right)(0)\beta=\pi_{\mathrm{G}}\beta$ so this transition occurs at the same rate as the transversion $C \to G$. As this is true for the other three kinds of transitions, there is no transition bias when $\kappa = 0$.

PAUP* will not let you specify κ, only the T-ratio. This is because the exact meaning of κ varies with different underlying base frequencies under different models. For example, the ratio of the rate of transitions to the rate of transversions for the HKY85 model differs depending on which transitions and transversions you compare. The ratio of the $A \to G$ transition to the $C \to G$ transversion is equal to κ under HKY85, but the ratio of the $A \to G$ transition to the $A \to C$ transversion is $\frac{\pi_G}{\pi_C}\kappa$. The T-ratio can be compared because its meaning is the same regardless of base frequencies.

Calculating the expected number of substitutions using matrix algebra

To calculate the probability of a site changing in a given instant of time (dt) we have to take into account the starting state of the site and determine the transition probability of that state in a given instant of time

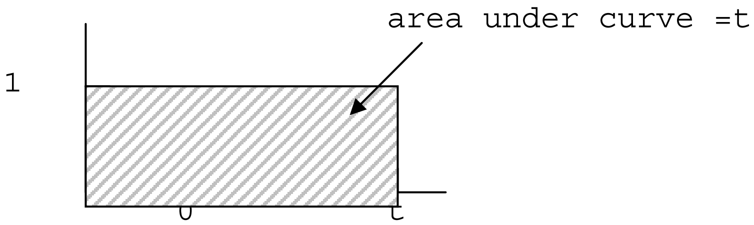
For example the probability of A→G substitution in an instant of time dt=

Pr(start with A) Pr(end with a G|dt) = $\left(\pi_{\mathrm{A}}\right)\left(\pi_{\mathrm{G}}\kappa\beta \mathrm{dt}\right)$

The time dt, is an infinitesimal amount of time during which only zero or one substitutions can occur. In this case, the expected number of substitutions equals the probability of a substitution. To obtain the expected number of substitutions over some arbitrarily large amount of time (t), we have to integrate the expected number of substitutions over dt from zero to t.

$$\int_0^{\mathrm{T}} \pi_{\mathrm{A}}\pi_{\mathrm{G}}\kappa\beta \mathrm{dt}$$

Move the first part out to leave dt $= \pi_{\mathrm{A}}\pi_{\mathrm{G}}\kappa\beta\int_0^{\mathrm{T}}(1)\mathrm{dt}$



So the probability of an A→G substitution occurring from time zero to t under model would be

$$= \pi_{\mathrm{A}}\pi_{\mathrm{G}}\kappa\beta t$$

This represents only one of probabilities in the Q-matrix. So in order to get the expected number of substitutions for the whole matrix we can use matrix algebra to multiply matrices together simplify the process. Recall the HKY85 Q-matrix from before

$$\mathrm{Q}=\begin{bmatrix} -\beta(\pi_{\mathrm{Y}}+\kappa\pi_{\mathrm{G}}) & \pi_{\mathrm{C}}\beta & \pi_{\mathrm{G}}\kappa\beta & \pi_{\mathrm{T}}\beta \\ \pi_{\mathrm{A}}\beta & -\beta(\pi_{\mathrm{R}}+\kappa\pi_{\mathrm{T}}) & \pi_{\mathrm{G}}\beta & \pi_{\mathrm{T}}\kappa\beta \\ \pi_{\mathrm{A}}\kappa\beta & \pi_{\mathrm{C}}\beta & -\beta(\pi_{\mathrm{Y}}+\kappa\pi_{\mathrm{A}}) & \pi_{\mathrm{T}}\beta \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\kappa\beta & \pi_{\mathrm{G}}\beta & -\beta(\pi_{\mathrm{R}}+\kappa\pi_{\mathrm{C}}) \end{bmatrix}$$

We can multiply the rate matrix Q by a time t to yield the following matrix:

$$\text{Qt}=\begin{bmatrix} -\beta(\pi_{\text{Y}}+\kappa\pi_{\text{G}})\text{t} & \pi_{\text{C}}\beta t & \pi_{\text{G}}\kappa\beta t & \pi_{\text{T}}\beta t \\ \pi_{\text{A}}\beta t & -\beta(\pi_{\text{R}}+\kappa\pi_{\text{T}})\text{t} & \pi_{\text{G}}\beta t & \pi_{\text{T}}\kappa\beta t \\ \pi_{\text{A}}\kappa\beta t & \pi_{\text{C}}\beta t & -\beta(\pi_{\text{Y}}+\kappa\pi_{\text{A}})\text{t} & \pi_{\text{T}}\beta t \\ \pi_{\text{A}}\beta t & \pi_{\text{C}}\kappa\beta t & \pi_{\text{G}}\beta t & -\beta(\pi_{\text{R}}+\kappa\pi_{\text{C}})\text{t} \end{bmatrix}$$

Remember to get the expected number of substitutions we have to multiply by the probability of the starting base (base freq). To do this for each term we can multiply the Π matrix by the Qt matrix to get:

$$\begin{bmatrix} \pi_{\text{A}} & 0 & 0 & 0 \\ 0 & \pi_{\text{C}} & 0 & 0 \\ 0 & 0 & \pi_{\text{G}} & 0 \\ 0 & 0 & 0 & \pi_{\text{T}} \end{bmatrix}\begin{bmatrix} -\beta(\pi_{\text{Y}}+\kappa\pi_{\text{G}})\text{t} & \pi_{\text{C}}\beta\text{t} & \pi_{\text{G}}\kappa\beta\text{t} & \pi_{\text{T}}\beta\text{t} \\ \pi_{\text{A}}\beta\text{t} & -\beta(\pi_{\text{R}}+\kappa\pi_{\text{T}})\text{t} & \pi_{\text{G}}\beta\text{t} & \pi_{\text{T}}\kappa\beta\text{t} \\ \pi_{\text{A}}\kappa\beta\text{t} & \pi_{\text{C}}\beta\text{t} & -\beta(\pi_{\text{Y}}+\kappa\pi_{\text{A}})\text{t} & \pi_{\text{T}}\beta\text{t} \\ \pi_{\text{A}}\beta\text{t} & \pi_{\text{C}}\kappa\beta\text{t} & \pi_{\text{G}}\beta\text{t} & -\beta(\pi_{\text{R}}+\kappa\pi_{\text{C}})\text{t} \end{bmatrix}$$

The rules for multiplying matrices demonstrated were

$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix}=\begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$ This is a simple example but if these rules are applied to the above matrices it yields

$$\Pi\text{Qt}=\begin{bmatrix} -\beta\text{t}\pi_{\text{A}}(\pi_{\text{Y}}+\kappa\pi_{\text{G}}) & \pi_{\text{A}}\pi_{\text{C}}\beta\text{t} & \pi_{\text{A}}\pi_{\text{G}}\kappa\beta\text{t} & \pi_{\text{A}}\pi_{\text{T}}\beta\text{t} \\ \pi_{\text{C}}\pi_{\text{A}}\beta\text{t} & -\beta\text{t}\pi_{\text{C}}(\pi_{\text{R}}+\kappa\pi_{\text{T}}) & \pi_{\text{C}}\pi_{\text{G}}\beta\text{t} & \pi_{\text{C}}\pi_{\text{T}}\kappa\beta\text{t} \\ \pi_{\text{G}}\pi_{\text{A}}\kappa\beta\text{t} & \pi_{\text{G}}\pi_{\text{C}}\beta\text{t} & -\beta\text{t}\pi_{\text{G}}(\pi_{\text{Y}}+\kappa\pi_{\text{A}}) & \pi_{\text{G}}\pi_{\text{T}}\beta\text{t} \\ \pi_{\text{T}}\pi_{\text{A}}\beta\text{t} & \pi_{\text{T}}\pi_{\text{C}}\kappa\beta\text{t} & \pi_{\text{T}}\pi_{\text{G}}\beta\text{t} & -\beta\text{t}\pi_{\text{T}}(\pi_{\text{R}}+\kappa\pi_{\text{C}}) \end{bmatrix}$$

We can now take the negative of the trace (add the diagonal terms and multiply by -1) of this matrix to solve for the total expected number of substitutions *d*

$$\begin{aligned} d &= \text{-trace}(\mathbf{\Pi Q}t) \\ &= \beta t[\pi_{\text{A}}(\pi_{\text{Y}}+\kappa\pi_{\text{G}})+\pi_{\text{C}}(\pi_{\text{R}}+\kappa\pi_{\text{T}})+\pi_{\text{G}}(\pi_{\text{Y}}+\kappa\pi_{\text{A}})+\pi_{\text{T}}(\pi_{\text{R}}+\kappa\pi_{\text{C}})] \\ &= \beta t[2\pi_{\text{A}}\pi_{\text{G}}\kappa+2\pi_{\text{C}}\pi_{\text{T}}\kappa+2\pi_{\text{R}}\pi_{\text{Y}}] \\ &= 2\beta t[\kappa(\pi_{\text{A}}\pi_{\text{G}}+\pi_{\text{C}}\pi_{\text{T}})+\pi_{\text{R}}\pi_{\text{Y}}] \end{aligned}$$

We can check this by seeing if the expected number of substitutions for the HKY85 are equivalent to that for the JC model if we plug in 1/4 base frequencies and equal rates.

For the JC model d=3αt, so when we set κ=1 and π_{base}=1/4

$$= 2\beta t\left[(1)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)+\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right] = 2\beta t\left[(1)\left(\frac{1}{16}\right)+\left(\frac{1}{16}\right)+\left(\frac{1}{4}\right)\right]$$

$$= (2)\left(\frac{3}{8}\right)\beta t = \left(\frac{3}{4}\right)\beta t = 3\alpha t$$

You can perform the algebra to see if this works for the K2P model as well.

Using matrix algegra allows more flexibility in the models that can be used. It is possible to do the matrix multiplication numerically, and so the transition probabilities needed for likelihood calculations can be obtained even for models in which it is impossible to find algebraic formulas for the transition probabilities.

Why the GTR model only has 5 relative rates

Recall the GTR matrix allows different rates of change for each nucleotide pair.

$$Q=\begin{bmatrix} - & \pi_C\beta a & \pi_G\beta b & \pi_T\beta c \\ \pi_A\beta a & - & \pi_G\beta d & \pi_T\beta e \\ \pi_A\kappa\beta b & \pi_C\beta d & - & \pi_T\beta f \\ \pi_A\beta c & \pi_C\kappa\beta e & \pi_G\beta f & - \end{bmatrix}$$

If base frequencies were equal, the expected number of substitutions for the matrix would be

$$d = \frac{1}{8}\beta t\left[a + b + c + d + e + f\right]$$

in the above equation, if we allow all of the relative rates (a, b, c, d, e, and f) to vary, then there are many possible combinations that will give the same value for d. For example, halving the value of β and doubling the value of each of the 6 relative rates would produce the same value for *d* as the above formula because the 2 can be factored out and cancels with the 2 in the denominator:

$d = \left(\frac{1}{8}\right)\left(\frac{\beta}{2}t\right)\left[2a + 2b + 2c + 2d + 2e + 2f\right]$ If one of the rates is

If one of the seven rate parameters {β, a, b, c, d, e, f} is constrained to equal 1, however, then d is uniquely defined by the parameters in the model. For example, setting f = 1 gives

$$d = \left(\frac{1}{8}\right)\left(\frac{\beta}{2}\mathrm{t}\right)\left[2\mathrm{a} + 2\mathrm{b} + 2\mathrm{c} + 2\mathrm{d} + 2\mathrm{e} + 1\right]$$

Here, we have again halved b and doubled all the relative rate parameters, but now the 2 cannot be factored out of the sum on the right side and thus halving β and doubling the relative rates changes the value of *d*, as it should.

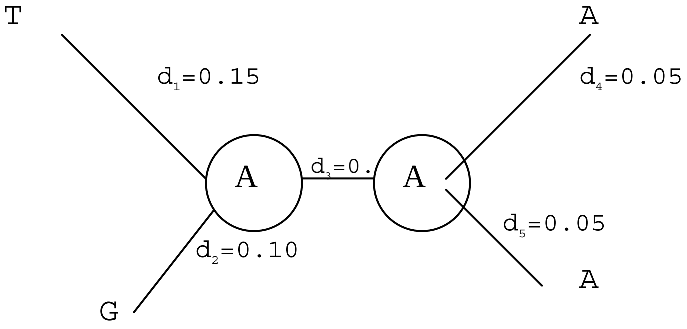
Calculating the likelihood of a 4 taxon tree

There is a worksheet associated with this.

We calculated the likelihood for one site, k, on a 4 taxon tree, producing a "site likelihood" L_k. Remember, to get the overall likelihood we would have to multiply this by the the site likelihoods for all other sites in the data matrix

$$L=L_1\ L_2\ L_3\ L_4\ L_5\ L_6\ L_7 \ldots\ldots\ldots\ L_k$$

To calculate the site likelihood assume branch lengths in the below figure equal the expected number of substitutions. Since the models are time reversible, it does not matter what point in the tree we consider the "root".



If we arbitrarily choose to start at G we calculate the site likelihood for this combination of internal node states

$$L_k = \mathrm{Pr(G)Pr(G \rightarrow A \mid d_2)Pr(A \rightarrow T \mid d_1)Pr(A \rightarrow A \mid d_3)Pr(A \rightarrow A \mid d_4)Pr(A \rightarrow A \mid d_5)}$$

Where Pr(G) is the probability of starting with base G and $\mathrm{Pr(G \rightarrow A \mid d_2)}$is the probability of changing from a G to an A along branch d_2 etc.

Using the K2P substitution model this becomes

$$L_k = \left(\frac{1}{4}\right)\left(\frac{1}{4}+\frac{1}{4}e^{-4(\beta t)_2} - \frac{1}{2}e^{-4(\beta t)_2\left(\frac{\kappa+1}{2}\right)}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4(\beta t)_1}\right)\left(\frac{1}{4}+\frac{1}{4}e^{-4(\beta t)_3} + \frac{1}{2}e^{-4(\beta t)_3\left(\frac{\kappa+1}{2}\right)}\right)$$

$$\cdot\left(\frac{1}{4}+\frac{1}{4}e^{-4(\beta t)_4} + \frac{1}{2}e^{-4(\beta t)_4\left(\frac{\kappa+1}{2}\right)}\right)\left(\frac{1}{4}+\frac{1}{4}e^{-4(\beta t)_5} + \frac{1}{2}e^{-4(\beta t)_5\left(\frac{\kappa+1}{2}\right)}\right)$$

The first term above is the probability of a transition, the second is the probability of a transversion, and the last three terms represent the probability of no change.

Pr(Transition) = $\frac{1}{4}+\frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-4\beta t\left(\frac{\kappa+1}{2}\right)}$

Pr(Transversion) = $\frac{1}{4} - \frac{1}{4}e^{-4\beta t}$

Pr(No Change) = $\frac{1}{4}+\frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-4\beta t\left(\frac{\kappa+1}{2}\right)}$

The below table summarizes the transition probabilities for the set of branch lengths on the example tree under the K2P model. We can calculate the quantity 4β*t* used in the transition formulas above given the branch length *d* because *d* = (κ+2)β*t* for the K80 model. For these calculations, we assumed *k* = 6, thus 4β*t* = 4*d* /(6+2) = *d*/2.

	d	4βt	Pr(no change)	Pr(Ti)	Pr(Ts)
d_3, d_4, d_5	0.05	0.025	0.951937	0.035718	0.00617252
d_2	0.1	0.05	0.9075	0.068079	0.0121926
d_1	0.15	0.075	0.866499	09	0.018064

Plugging the values from the table above into the equation below:

$$L_k = \Pr(G)\Pr(G \to A \mid d_2)\Pr(A \to T \mid d_1)\Pr(A \to A \mid d_3)\Pr(A \to A \mid d_4)\Pr(A \to A \mid d_5)$$

we get

$$L_k = \left(\frac{1}{4}\right)(0.068079)(0.0180641)(0.951937)^3 = .00026521$$

This is the likelihood for one site with one combination of ancestral states and one particular combination of branch lengths and κ. In order to get the total likelihood, we have to calculate this value for all possible combinations of ancestral states. For a 4 taxon tree this means we must calculate it for all 16 possible combinations of ancestral states and then add these together.

We did this in class for the above example, Paul provided the table below for the notes

kappa = 6

	d3,d4,d5	d2	d1
	0.05	0.10	0.15
Pr(same)	0.951936914	0.907535867	0.866499054
Pr(trs)	0.035718042	0.068078846	0.097372689
Pr(trv)	0.006172522	0.012192644	0.018064128

```
Anc.                Pct.
State   likelihood  of sum
--------------------------
AA      0.000265212  63.2%
AC      7.23031E-11   0.0%
AG      1.40098E-08   0.0%
AT      7.23031E-11   0.0%
CA      1.66018E-06   0.4%
CC      1.07649E-08   0.0%
CG      2.33729E-09   0.0%
CT      4.03913E-10   0.0%
GA      0.000132655  31.6%
GC      9.63848E-10   0.0%
GG      4.97742E-06   1.2%
GT      9.63848E-10   0.0%
TA      1.47736E-05   3.5%
TC      3.59434E-09   0.0%
TG      2.07991E-08   0.0%
TT      9.57943E-08   0.0%
         0.000419429   100% = total site likelihood
```

Note added by Paul:

The NEXUS file used to calculate these likelihoods in PAUP* is also below. This NEXUS file contains 4 blocks: a PAUP block, a CHARACTERS block, a TREES block, and another PAUP block.

The first PAUP block sets the criterion to maximum likelihood (crit=like) and tells PAUP* to store any branch lengths it finds (storebrlens). The second line sets up PAUP* to use the K80 model: nst=2 means we want to use a model with two substitution classes (transitions and ransversions), tratio=3.0 sets =6 (because tratio= /2 for this model), basefreq=equal means all relative base frequencies are to be ¼, and variant=hky is necessary to distinguish this model (which is identical to the hky model with equal frequencies) from the F84 model with equal frequencies (if you wanted F84, you would specify variant=f84). Note: the main reason for having two PAUP blocks is to get storebrlens specified before PAUP* reads in the TREES block. The other stuff could be moved to the second PAUP block.

The CHARACTERS block provides a small data matrix containing 2 sites for 4 taxa. The CHARACTERS block like a DATA block, with which you are probably more familiar.

The TREES block provides the tree description. Note that taxon5 is listed in the tree description, but there is no data for taxon5 in the data matrix. taxon5 is simply being used as a label for one of the interior nodes of the tree here. The utree designation tells paup that this is an unrooted tree (there is no meaningful root to this tree).

The final PAUP block tells paup to compute the likelihood scores (lscores) for individual sites (sitelikes), using the branch lengths (userbrlens) that we provided (i.e. don't try to estimate branch lengths using maximum likelihood).

```
#nexus

begin paup;
  set crit=like storebrlens;
  lset nst=2 variant=hky tratio=3.0 basefreq=equal;
end;

begin characters;
  dimensions newtaxa ntax=4 nchar=2;
  format datatype=dna;
  matrix
    taxon1 TA
    taxon2 GC
    taxon3 AG
    taxon4 AT
  ;
end;

begin trees;
  utree test = ((taxon1:0.15,taxon2:0.10)taxon5:0.05,taxon3:0.05,taxon4:0.05);
end;

begin paup;
  lscores 1 / userbrlens sitelikes;
end;
```