Phylogenetics

Andreas Bernauer, andreas@carrot.mcb.uconn.edu

March 28, 2004

Contents

1 ts:tr rate ratio vs. ts:tr ratio **1**

2 Expected number of substitutions using matrix algebra **2**

3 Why the GTR model can only have 5 relative rates **6**

4 Likelihood of a 4 taxa tree **6**
4.1 Note added by Paul: 8

1 ts:tr rate ratio vs. ts:tr ratio

The transition:transversion (ts:tr) *rate* ratio (called `kappa` in PAUP) is different from the ts:tr ratio (called `tratio` in PAUP). Let's recall the K80 model (also called K2P model) and look what these two ratios look like:

$$Q = \begin{bmatrix} -\beta(\kappa+2) & \beta & \kappa\beta & \beta \\ \beta & -\beta(\kappa+2) & \beta & \kappa\beta \\ \kappa\beta & \beta & -\beta(\kappa+2) & \beta \\ \beta & \kappa\beta & \beta & -\beta(\kappa+2) \end{bmatrix}$$

The ts:tr rate ratio (`kappa`)is the ratio of the transition rate and the transversion rate. In the K80 model the transition rate is $\kappa\beta$ and the transversion rate is β. Thus, the ts:tr rate ratio of the K80 model is κ:

$$\frac{\text{transition rate}}{\text{transversion rate}} = \frac{\kappa\beta}{\beta} = \kappa$$

The ts:tr ratio (`tratio`) is the ratio of two probabilities: the probability of any transition over a time dt and the probability of any transversion over the same time dt. Both probabilities are the sum of the probabilities of individual transitions or transversions, respectively. For example, the probability of the transition $A \to G$ is the probability of having an A, times the probability of an

A changing to a G over time dt. The ts:tr ratio for the K80 model is calculated like this:[1]

$$\begin{aligned}
\texttt{tratio} &= \frac{\text{Prob\{any transition over time dt\}}}{\text{Prob\{any transversion over time dt\}}} \\
&= \frac{\frac{1}{4}\,\mathrm{P}_{AG}(dt) + \frac{1}{4}\,\mathrm{P}_{CT}(dt) + \frac{1}{4}\,\mathrm{P}_{GA}(dt) + \frac{1}{4}\,\mathrm{P}_{TC}(dt)}{\frac{1}{4}\,\mathrm{P}_{AC}(dt) + \frac{1}{4}\,\mathrm{P}_{CA}(dt) + \cdots + \frac{1}{4}\,\mathrm{P}_{GT}(dt) + \frac{1}{4}\,\mathrm{P}_{TG}(dt)} \\
&= \frac{4(\frac{1}{4}\kappa\beta dt)}{8(\frac{1}{4}\beta dt)} \\
&= \frac{\kappa}{2}
\end{aligned}$$

In a model that incorporates base frequencies the ratios are calculated in the same way, but yield more complicated results. Let's take the HKY85 model as an example (diagonal elements omitted):

$$Q = \begin{bmatrix} - & \pi_C\beta & \pi_G\kappa\beta & \pi_T\beta \\ \pi_A\beta & - & \pi_G\beta & \pi_T\kappa\beta \\ \pi_A\kappa\beta & \pi_C\beta & - & \pi_T\beta \\ \pi_A\beta & \pi_C\kappa\beta & \pi_G\beta & - \end{bmatrix}$$

The tr:ts rate ratio (called `kappa` in PAUP) may now also depend on the base frequencies, if the transition and the transversion are picked from different columns. Thus, there is no single tr:ts rate ratio for this model. However, there is a single tr:ts ratio (called `tratio` in PAUP), which is a function of the base frequencies, π_i, κ and β. This is the reason why PAUP concentrates on `tratio`: this quantity has a single meaning regardless of nucleotide composition and it is therefore comparable among different models.

A side note on maximum parsimony: maximum parsimony does not use any models presented here, but does allow for weighting transitions and transversions differently.

Note added by Paul: In parsimony the weights have no particular meaning. You can give transversions a weight 5 times that of transitions, but this value (5) does not correspond to a ts: tr ratio of 5, or a ts:tr rate ratio of 5. Unlike maximum likelihood, maximum parsimony provides no means for determining what weights are appropriate.

2 Expected number of substitutions using matrix algebra

In this section we review how to calculate the expected number of substitutions and see how we can use matrix algebra for this purpose. Using matrix algebra

[1] The denominator contains all 8 transversions: AC, CA, AT, TA, CG, GC, GT, and TG

has the advantage that we can easily use mathematical programs to do the calculations for us.

First, let's see how we can get the expected number of substitutions for a small period of time, dt, based on the HKY85 model. As a shorthand, let's let π_Y be the relative frequency of pyrimidines and π_R be the relative frequency of purines:

$$\begin{aligned}\pi_Y &= \pi_C + \pi_T \\ \pi_R &= \pi_A + \pi_G\end{aligned}$$

With that, the Q matrix of the HKY85 model looks like this:

$$Q = \begin{bmatrix} -\beta(\pi_Y + \kappa\pi_G) & \pi_C\beta & \pi_G\kappa\beta & \pi_T\beta \\ \pi_A\beta & -\beta(\pi_R + \kappa\pi_T) & \pi_G\beta & \pi_T\kappa\beta \\ \pi_A\kappa\beta & \pi_C\beta & -\beta(\pi_Y + \kappa\pi_A) & \pi_T\beta \\ \pi_A\beta & \pi_C\kappa\beta & \pi_G\beta & -\beta(\pi_R + \kappa\pi_C) \end{bmatrix}$$

We know from previous lectures that the expected value of some (discrete) distribution Y is the sum of the each possible value for Y times the probability of this value:

$$E(Y) = \sum_i y_i \,\mathrm{P}(y_i)$$

In our small amount of time dt, there can be only 0 or 1 substitutions. Thus, the expected value for our distribution breaks down to:

$$\begin{aligned}E(Y) &= (0)\,\mathrm{P}(0) + (1)\,\mathrm{P}(1) \\ &= \mathrm{P}(1)\end{aligned}$$

This means, if we want to know the expected number of substitutions over time dt for, *e.g.* $A \to G$ substitutions, we only have to calculate the probability for this event:

$$\begin{aligned}\text{Exp. \# of } A \to G \text{ subst. in } dt &= \mathrm{P}(A) \cdot \mathrm{P}(A \to G | dt) \\ &= \pi_A \cdot \pi_G\kappa\beta\, dt \end{aligned} \tag{1}$$

If we want to know the expected number of substitutions over an arbitrary time interval t, we have to take the continuous sum, *i.e.* the integral over this time interval:

$$\begin{aligned}\text{Exp. \# of } A \to G \text{ subst. in } t &= \int_0^t \pi_A\pi_G\kappa\beta\, dt \\ &= \pi_A\pi_G\kappa\beta \int_0^t 1\, dt \\ &= \pi_A\pi_G\kappa\beta t \end{aligned} \tag{2}$$

We see, that we just have substituted dt by t from formula (1) to formula (2). This is valid here (as we have shown by calculation), but might not be valid in other cases.

If we want to get the overall number of expected substitutions (not only the number for $A \to G$), we have to sum up the number of expected substitutions for all 12 possibilities. We do this by either performing the calculations shown above 12 times, or by using matrix algebra and doing it for all 12 possibilities at once. For this, we need to transform each entry in the Q matrix of the HKY85 model into the form we have derived above. Thus, we need to perform the following steps:

- Multiply the Q matrix by t, getting Qt:

$$Qt = \begin{bmatrix} -\beta(\pi_Y + \kappa\pi_G)t & \pi_C\beta t & \pi_G\kappa\beta t & \pi_T\beta t \\ \pi_A\beta t & -\beta(\pi_R + \kappa\pi_T)t & \pi_G\beta t & \pi_T\kappa\beta t \\ \pi_A\kappa\beta t & \pi_C\beta t & -\beta(\pi_Y + \kappa\pi_A)t & \pi_T\beta t \\ \pi_A\beta t & \pi_C\kappa\beta t & \pi_G\beta t & -\beta(\pi_R + \kappa\pi_C)t \end{bmatrix}$$

- Multiply each row of Qt by the appropriate base frequency. We can do this by performing a left side matrix multiplication of Qt with Π, the matrix that has the base frequency in its diagonal and zero anywhere else:

$$\Pi Qt = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix} \begin{bmatrix} -\beta(\pi_Y + \kappa\pi_G)t & \pi_C\beta t & \pi_G\kappa\beta t & \pi_T\beta t \\ \pi_A\beta t & -\beta(\pi_R + \kappa\pi_T)t & \pi_G\beta t & \pi_T\kappa\beta t \\ \pi_A\kappa\beta t & \pi_C\beta t & -\beta(\pi_Y + \kappa\pi_A)t & \pi_T\beta t \\ \pi_A\beta t & \pi_C\kappa\beta t & \pi_G\beta t & -\beta(\pi_R + \kappa\pi_C)t \end{bmatrix}$$

$$= \begin{bmatrix} -\beta\pi_A(\pi_Y + \kappa\pi_G)t & \pi_A\pi_C\beta t & \pi_A\pi_G\kappa\beta t & \pi_A\pi_T\beta t \\ \pi_C\pi_A\beta t & -\beta\pi_C(\pi_R + \kappa\pi_T)t & \pi_C\pi_G\beta t & \pi_C\pi_T\kappa\beta t \\ \pi_G\pi_A\kappa\beta t & \pi_G\pi_C\beta t & -\beta\pi_G(\pi_Y + \kappa\pi_A)t & \pi_G\pi_T\beta t \\ \pi_T\pi_A\beta t & \pi_T\pi_C\kappa\beta t & \pi_T\pi_G\beta t & -\beta\pi_T(\pi_R + \kappa\pi_C)t \end{bmatrix}$$

- Sum all 12 off-diagonal elements. This can be tedious in a program for matrix calculation as we have to write down a sum of 12 elements. For amino acid matrices this is even more tedious, as the sum consists of 380 elements.

  We can use a shorthand for that, as every diagonal element is the negative sum of the off-diagonal elements of its row. Thus, if we add all diagonal elements and negate the result, we get the same as if we had summed all off-diagonal elements. Adding all diagonal elements is done by the matrix

function trace:

$$
\begin{aligned}
d &= \text{Exp. overall \# of substitutions over time } t \\
&= -\operatorname{trace}(\Pi Q t) \\
&= \beta t[\pi_A\pi_Y + \pi_A\pi_G\kappa + \pi_C\pi_R + \pi_C\pi_T\kappa + \pi_G\pi_Y + \pi_G\pi_A\kappa + \pi_T\pi_R + \pi_T\pi_C\kappa] \\
&= \beta t[2\pi_R\pi_Y + 2\pi_A\pi_G\kappa + 2\pi_C\pi_T\kappa] \\
&= 2\beta t[\pi_R\pi_Y + \kappa(\pi_A\pi_G + \pi_C\pi_T)] \qquad (3)
\end{aligned}
$$

Let's compare the result of (3) with what we know about the expected number of substitutions in the JC model and the K80 model:

$$
\begin{aligned}
d_{JC} &= 3\alpha t \\
d_{K80} &= (\kappa + 2)\beta t
\end{aligned}
$$

In the JC model, $\kappa = 1$ and the base frequencies $\pi_i = 1/4$. Using formula (3), we get:

$$
\begin{aligned}
d_{JC} &= 2\beta t[\frac{1}{2}\frac{1}{2} + 1 \cdot (\frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{1}{4})] \\
&= 2\beta t\frac{3}{8} \\
&= \frac{3}{4}\beta t \\
&= 3\alpha t \qquad \text{as } \alpha = \frac{1}{4}\beta
\end{aligned}
$$

which is what we expected.

In the K80 model, the base frequencies are 1/4 as in the JC model. The β of the K80 model, however, is $\frac{1}{4}\beta$ in the HKY85 model. In the following, I say β_{K80} for the β of the K80 model. Plugging in the values in formula (3) results in:

$$
\begin{aligned}
d_{K80} &= 2\beta t[\frac{1}{2}\frac{1}{2} + \kappa \cdot (\frac{1}{4}\frac{1}{4} + \frac{1}{4}\frac{1}{4})] \\
&= \frac{1}{2}\beta + \beta t\frac{1}{4}\kappa \\
&= \frac{1}{4}\beta t(2 + \kappa) \\
&= \beta_{K80} t(2 + \kappa)
\end{aligned}
$$

which is again what we expected.

3 Why the GTR model can only have 5 relative rates

Recall the GTR model of last class (diagonal elements omitted):

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

In the last class we stated that this model has the following parameters: The three base frequencies π_A, π_C, π_G (the fourth frequency is given by the fact that the four frequencies sum up to 1), β and five relative rates a, b, c, d, e. The question was, why only five of the six relative rates can be considered as parameters. There was no easy reasoning for this as for the base frequencies.

Let's see what happens if all six relative rates a, ..., f were parameters of the GTR model. When we look at the expected number of substitutions d, and assume for simplicity that the base frequencies are all $\frac{1}{4}$, we get:

$$d = \frac{1}{8}\beta t(a + b + c + d + e + f)$$

which is the same as

$$= \frac{1}{8}\frac{\beta}{2}t(2a + 2b + 2c + 2d + 2e + 2f)$$

Thus, two different set of parameters gave the same d: β, a, b, c, d, e, f as well as $\frac{\beta}{2}$, $2a$, $2b$, $2c$, $2d$, $2e$, $2f$. Actually there were an infinite number of parameter sets that gave the same value for d, as we can divide or multiply by any other number than 2. This is unfortunate, as this reduces the usability of our model: if you were to compare the results for two datasets by comparing the values for βt, you couldn't do it without looking at the other six parameters a, ..., f, as the value for βt depends on these parameters. However, if there are only five relative rate parameters, you cannot do the above manipulation anymore, as you can't freely chose one of the relative rates. Thus, there is only one possible value for βt and you are able to compare the results.

Note, that there is no real reason why β is a parameter, but f is not. You could choose f to be a parameter, but β not to be. It has only historic / traditional reasons why β is considered a parameter.

4 Likelihood of a 4 taxa tree

In this section we want to calculate the likelihood of a 4 taxa tree. If n is the length of the taxa sequences, then the likelihood L of the tree is:

$$L = L_1 L_2 \cdots L_k \cdots L_n$$

As the calculations for the whole sequence are elaborate, we only calculate the likelihood for one site of the sequence, L_k. In the class we got a worksheet that depicted a 4 taxa tree. The tree showed only the sites at position k of the taxa sequences, for which we are going to calculate the likelihood: ((T, G), (A, A)), with the root arbitrarily chosen to be at G. We don't know the states at the inner nodes, but we do know the distances: $d_1 = 0.15$ is the distance from T to the first inner node, $d_2 = 0.10$ is the distance from G to the first inner node, and $d_3 = d_4 = d_5 = 0.05$ are the remaining distances.

As we don't know the inner node states, L_k is the sum of the probabilities for all 16 possibilities at the inner nodes:

$$L_k = L_k^{AA} + \cdots + L_k^{AT} + L_k^{CA} + \cdots + L_k^{CT} + \cdots + L_k^{TT}$$

We assume a K80 model with parameter $\kappa = 6$. In the K80 model, the substitution probabilities are as follows:

$$\begin{aligned}
\mathrm{P}(\text{no change}) &= \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-4\beta t\left(\frac{\kappa+1}{2}\right)} \\
\mathrm{P}(\text{transition}) &= \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-4\beta t\left(\frac{\kappa+1}{2}\right)} \\
\mathrm{P}(\text{transversion}) &= \frac{1}{4} - \frac{1}{4}e^{-4\beta t}
\end{aligned}$$

Given κ and a distance $d = d_{K80}$, we can calculate the probabilities for no change, a transition, and a transversion under the K80 model. For this, we have to calculate $4\beta t$ which we can derive from a given distance:

$$\begin{aligned}
& d = (\kappa + 2)\beta t \\
\Leftrightarrow \quad & \beta t = \frac{d}{\kappa + 2} \\
\Leftrightarrow \quad & 4\beta t = \frac{4d}{\kappa + 2} \overset{\kappa=6}{=} \frac{d}{2}
\end{aligned}$$

Thus, the three possible probabilities for each of the distances are:

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

With these values we can calculate the likelihoods for each of the 16 possibilities for the inner nodes. *E.g.* if the inner nodes were both A, the likelihood is:

$$\begin{aligned}
L_k^{AA} &= \mathrm{P}(G)\,\mathrm{P}_{GA}(d_2)\,\mathrm{P}_{AT}(d_1)\,\mathrm{P}_{AA}(d_3)\,\mathrm{P}_{AA}(d_4)\,\mathrm{P}_{AA}(d_5) \\
&= \frac{1}{4} \cdot 0.068078846 \cdot 0.018064128 \cdot (0.951936914)^3 \\
&= 0.000265212
\end{aligned}$$

nodes ij	L_k^{ij}	percent of L_k
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%
$L_k =$	0.000419429	100.0%

Table 1: Likelihoods L_k^{ij} for each possibility of inner nodes for the tree described in the text and an assumed K80 model with $\kappa = 6$. The total site likelihood L_k is also shown, as well the fraction of each of the 16 component terms in the total likelihood.

We multiply by P(G) as this is the root. With the same approach we get the values for all 16 possible inner node combinations. Table 1 shows the values.

When we run the following PAUP program shown in figure 1, we get the same results.

4.1 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 , 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

```
#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;
```

Figure 1: PAUP program that does the same likelihood calculation as shown in the text.

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).