Phylomath Lecture 7
Other Substitution Models: K2P, HKY, F84, and GTR
Calculating Expected Number of Substitutions

Maxi Polihronakis (16 March 2004)

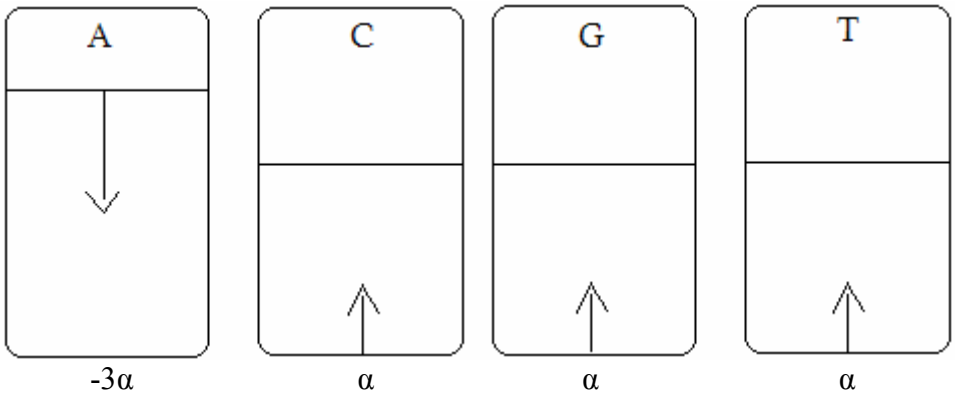
Substitution models can be viewed as instantaneous rate matrices (Q), which describe the instantaneous rate of substitution for all possible substitutions. We will begin with the Jukes-Cantor[1] matrix and use this to develop rate matrices for the Kimura 2 Parameter model (K2P or K80)[2], Felsenstein 1981 model (F81)[3], Hasegawa, Kishino, and Yano model (HKY)[4], Felsenstein 1984 model (but published by Kishino and Hasegawa in 1989[5]), and the General Time Reversible model (GTR)[6,7,8,9,10,11].

The instantaneous rate matrix of the J-C model is as follows:

Q =

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

This model assumes $\pi_A = \pi_C = \pi_G = \pi_T = ¼$ where π is used to designate a proportion (not 3.14159265358979323846264338327 95), and that transitions and transversions occur at equal rates. One important thing to keep in mind is that all rows of an instantaneous rate matrix must sum to zero. This can be explained using the analogy of four canisters filled with water representing the base pairs in a sequence.



Because there are a fixed number of sites in a sequence, when some sites change from A to say T, the level in canister T rises and the level in canister A falls by the same amount. Canister A is decreasing at rate 3α, while C, G, and T are increasing at rate α .

The J-C model is a simple model with relatively unrealistic assumptions. Additional models have been developed which allow us to incorporate factors we know affect molecular sequence evolution. The K80 model is different from the J-C model in that it does not assume transitions and transversions occur at the same rate. Although there are twice as many combinations of transversions (A←→C/A←→T/G←→C/G←→T) as transitions (A←→G/C←→T), transitions occur more often because of the chemical similarity within the purine and pyrimidine groups.

The instantaneous rate matrix for the K80 model is as follows:

$Q =$

	A	C	G	T
A	$-\beta(\kappa + 2)$	β	$\kappa\beta$	β
C	β	$-\beta(\kappa + 2)$	β	$\kappa\beta$
G	$\kappa\beta$	β	$-\beta(\kappa + 2)$	β
T	β	$\kappa\beta$	β	$-\beta(\kappa + 2)$

The diagonal spaces (A → A/C → C/G → G/T → T) are determined by subtracting the other terms in the row, for example:

$$-\beta - \kappa\beta - \beta$$
$$= \quad -\beta(\kappa + 2)$$

We have also introduced a new parameter, κ, which is known as the transition/transversion rate ratio. This conversion factor allows us to switch between transition and transversion rates. As the models become more complex, it is possible to constrain the additional parameters to return to a less complex model. For example, if we set $\kappa = 1$, the rate matrix for the K80 model will look the same as the J-C model. Thus, more complex models subtract assumptions by allowing for more flexibility.

Felsenstein 1981 (F81) is a model differing from the J-C model by allowing us to incorporate unequal base frequencies, while transition and transversion rates remain equal.

The instantaneous rate matrix for the F81 model is as follows:

$Q =$

	A	C	G	T
A	$-\beta(1 - \pi_A)$	$\pi_C\beta$	$\pi_G\beta$	$\pi_T\beta$
C	$\pi_A\beta$	$-\beta(1 - \pi_C)$	$\pi_G\beta$	$\pi_T\beta$
G	$\pi_A\beta$	$\pi_C\beta$	$-\beta(1 - \pi_G)$	$\pi_T\beta$
T	$\pi_A\beta$	$\pi_C\beta$	$\pi_G\beta$	$-\beta(1 - \pi_T)$

The diagonal is determined the same way as before by subtracting all of the off-diagonal terms from the row. The first row is simplified as follows:

$$- \pi_C\beta - \pi_G\beta - \pi_T\beta$$
$$= - \beta(\pi_C - \pi_G - \pi_T)$$

Because the proportion of all sites must add up to one, $\pi_C - \pi_G - \pi_T = 1 - \pi_A$, then

$$= - \beta(1 - \pi_A)$$

This rate matrix can look the same as the J-C matrix if we replace all values of π with ¼, then ¼ $\beta = \alpha$ or $\alpha = 4\beta$.

This model can be used to easily demonstrate the time reversibility property of all of the models discussed here. That is, the probability of starting with base i and then changing from to base j is the same as the probability of starting with base j and then changing to base i, or:

Pr (starting with i) Pr (changing to j) = Pr (starting with j) Pr (changing to i)

i ←————dt————→ j

We can see this is true by plugging in values from the F81 rate matrix and checking whether A ←—dt—→ C:

$$\Pr(A) \Pr(A \rightarrow C) = \Pr(C) \Pr(C \rightarrow A)$$
$$\pi_A(\pi_C\beta\, dt) = \pi_C(\pi_A\beta\, dt)$$

Since you are multiplying the same terms on each side of the equation, the statement is true.

Model	Parameters	#
J-C:	α	1
K80:	κ, β	2
F81:	β, π_A, π_C, π_G	4

So far, the complexity of our models has been increasing as we add parameters.

We only count the base frequency proportions in the F81 model as three free parameters because they must add up to one. If we know all three frequencies the fourth one is fixed.

The next two models, HKY85 and F84, are different from those previously mentioned because they allow for unequal base frequencies, as well as different transition and transversion rates. Although both accomplish similar goals, they do this in conceptually different ways.

	A	C	G	T
A	$-\beta(\pi_Y + \kappa\,\pi_G)$	$\pi_C\beta$	$\pi_G\beta\,\kappa$	$\pi_T\beta$
C	$\pi_A\beta$	$-\beta(\pi_R + \kappa\,\pi_T)$	$\pi_G\beta$	$\pi_T\beta\,\kappa$
G	$\pi_A\beta\,\kappa$	$\pi_C\beta$	$-\beta(\pi_Y + \kappa\,\pi_A)$	$\pi_T\beta$
T	$\pi_A\beta$	$\pi_C\beta\,\kappa$	$\pi_G\beta$	$-\beta(\pi_R + \kappa\,\pi_C)$

The instantaneous rate matrix for the HKY85 model is as follows:

Where $\pi_Y = \pi_C + \pi_T$ and $\pi_R = \pi_A + \pi_G$

In order to get back to the F81 model, we just set $\kappa = 1$. To get back to the K80 model, we set $\pi_C + \pi_G + \pi_T + \pi_A = ¼$. If we combine these two we return back to the J-C model. This model adds an additional parameter, κ, to the F81 model, and has a total of five free parameters. This model essentially combines the parameters from the K80 and F81 models in order to allow for both factors to vary. The F84 model, on the other hand, takes two processes into consideration when determining the rate matrix. The first process is called the *general* substitution process, which can generate any type of substitution and corresponds to the β rate parameter in the above models. The second process is called the *within-group* substitution process, and only generates transitions (i.e. purine-to-purine substitutions or pyrimidine-to-pyrimidine substitutions). The relative frequency that is used for this within-group process is the frequency of the base we are switching to divided by the others in the same group. For example, a substitution from a purine to a purine would have the rate: $(\pi_A/\ \pi_A\ +\ \pi_G)\ \Phi\beta$. The within-group

substitution process thus occurs at a rate F times the rate of the general substitution process.

The instantaneous rate matrix for the F84 model is as follows:

	A	C	G	T
A	$-\beta(1-\pi_A)$ $-\Phi\beta(\pi_G/\pi_R)$	$\pi_C\beta$	$\pi_G\beta$ $+(\pi_G/\pi_A+\pi_G)\Phi\beta$	$\pi_T\beta$
C	$\pi_A\beta$	$-\beta(1-\pi_C)$ $-\Phi\beta(\pi_T/\pi_Y)$	$\pi_G\beta$	$\pi_T\beta$ $+(\pi_T/\pi_C+\pi_T)\Phi\beta$
G	$\pi_A\beta$ $+(\pi_A/\pi_A+\pi_G)\Phi\beta$	$\pi_C\beta$	$-\beta(1-\pi_G)$ $-\Phi\beta(\pi_A/\pi_R)$	$\pi_T\beta$
T	$\pi_A\beta$	$\pi_C\beta$ $+(\pi_C/\pi_C+\pi_T)\Phi\beta$	$\pi_G\beta$	$-\beta(1-\pi_T)$ $-\Phi\beta(\pi_C/\pi_Y)$

Where $\pi_Y = \pi_C + \pi_T$ and $\pi_R = \pi_A + \pi_G$. One can notice from this matrix that if $\Phi > 0$, then transitions occur at a higher average rate than transversions. In other words, instead of using the factor κ to alter transversion rates, this model uses a different concept of within group substitutions, but maintains the same number of parameters. One minor advantage to this model over the HKY85 model, is that you can write out the formulas for the transition probabilities, making it possible write out the likelihood as an expression. A question was raised regarding the biological reality of this model versus the HKY85 model. The bottom line: no model is an accurate representation of reality, however the F84 model gives similar results to the HKY85 model.

To get the F81 model from the F84 model, set $\Phi = 0$. This is different than the HKY85 model where we set $\kappa = 1$ to go back to the F81 model. To get the K80 model from the F84 model you will need to set $\pi_i = ¼$. In addition, it is necessary to find a value for Φ such that all transitions have a rate equal to $1/4\beta\,\kappa$ when $\pi_i = ¼$. Working from the AG cell of the rate matrix:

$$\pi_G\beta + (\pi_G/\pi_R)\Phi\beta = \pi_G\beta\,\kappa$$
$$¼\beta + ½\Phi\beta = ¼\beta\,\kappa$$
$$1 + 2\Phi = \kappa$$
$$\Phi = ½(\kappa - 1)$$

If we plug this in we can check our work:

$$¼\,\beta + ½\Phi\beta$$
$$= ¼\,\beta + ½[½(\kappa-1)]\beta$$
$$= ¼\,\beta + ¼(\kappa-1)\beta$$

$= \frac{1}{4}\beta + \frac{1}{4}\kappa\beta - \frac{1}{4}\beta$
$= \frac{1}{4}\beta\kappa$

To get the J-C model: $\pi_i = \frac{1}{4}$ and $\Phi = 0$.

The last model we encountered was the GTR model. This model allows six distinct substitution types to each have their own rate.

The instantaneous rate matrix for the GTR model is as follows:

	A	C	G	T
A	$-\beta(\pi_C a + \pi_G b + \pi_T c)$	$\pi_C \beta a$	$\pi_G \beta b$	$\pi_T \beta c$
C	$\pi_A \beta a$	$-\beta(\pi_A a + \pi_G d + \pi_T e)$	$\pi_G \beta d$	$\pi_T \beta e$
G	$\pi_A \beta b$	$\pi_C \beta d$	$-\beta(\pi_A b + \pi_C d + \pi_T f)$	$\pi_T \beta f$
T	$\pi_A \beta c$	$\pi_C \beta e$	$\pi_G \beta f$	$-\beta(\pi_A c + \pi_C e + \pi_G f)$

To return back to the J-C model set:

$\pi_i = \frac{1}{4}$ and
a,b,c,d,e,f all = 1

To get back to the K80 model set:

$\pi_i = \frac{1}{4}$ and
a,c,d,f all = 1
$b = e = \kappa$

The number of free parameters when using the GTR model is 9:

β, π_A, π_C, π_G, a, b, c, d, e

Paul will reveal next week why the last parameter (f) is not free to vary.

We can use the ML calculator computer program to play around with four of the models (all except GTR and F84) to experience how difficult it is to estimate the parameters. Each free parameter in the model chosen can be increased or decreased simultaneously in order to find the right combinations so that the likelihood is maximized. This is similar to how programs like PAUP* work when estimating parameters for the chosen model of molecular sequence evolution.

The last part of lecture was dedicated to learning how to estimate the total number of substitutions for a sequence. This was demonstrated using the K80 rate matrix:

Q =

	A	C	G	T
A	$\beta(\kappa + 2)$	β	$\kappa\beta$	β
C	β	$\beta(\kappa + 2)$	β	$\kappa\beta$
G	$\kappa\beta$	β	$\beta(\kappa + 2)$	β
T	β	$\kappa\beta$	β	$\beta(\kappa + 2)$

We will first begin by asking how many A → G substitution we would expect over a small period of time dt:

We start with the equation:

Pr (A) Pr (A → G) and plug in values from the rate matrix
= (¼) ($\kappa\beta$ dt)

Multiplying by dt, allows us to turn the rate into a probability (more precisely, it turns the rate into the expected number of substitutions over the dt time interval, but this is equal to the probability of a substitution because there can be only 0 or 1 substitutions in the time dt). If we do this for all types of substitutions using the matrix above and applying matrix algebra, we will be all set. Thus each substitution rate in the matrix is multiplied by ¼ and dt and summed together. Rather than summing 12 quantities, it is easier to just sum the 4 terms on the main diagonal. This sum of terms along the main diagonal is known as the *trace* of the matrix to professionals. Because each of these diagonal elements is the negative sum of the other elements on the same row, computing the trace is equivalent to summing all 12 off-diagonal elements.

$$- \text{trace}\,(1/4\ Q\ dt) = 4\ (1/4)\ (\kappa + 2)\ \beta\ dt$$
$$= (\kappa + 2)\ \beta\ dt$$

In order to sum $(\kappa + 2)\beta dt$ over the infinite number of small intervals between 0 and t, we integrate:

$$\int_0^t (\kappa + 2)\beta dt$$
$$= (\kappa + 2)\beta \int_0^t dt$$
$$= (\kappa + 2)\beta\, t$$

Thus, the expected number of substitutions = $(\kappa + 2)\, \beta\, t$

If we use the HKY85 model instead of the K80 model, we have π_i instead of ¼. Therefore, instead of multiplying Q by ¼, we multiply each row of the matrix by its respective base frequency because this is the probability of starting with that base.

π_A across the first row,
π_C across the second row,
π_G across the third row, and
π_T across the fourth row.

Take the negative trace of the matrix and you will arrive at the expected number of substitutions, try it and see.

The purpose of estimating the total number of substitutions in a sequence is to have a standard way of describing how much evolution has occurred. This is one way of combining all of the parameters from the different models in order to make comparisons.

References

1. Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules.
Pages 21-132 in Mammalian Protein Metabolism (H. N. Munro, ed.) Academic Press, New York.

2. Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111-120.

3. Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368-376.

4. Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 21:160-174.

5. Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution 29:170-179.

6. Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86-93.

7. Rodriguez, F., O. J. L., A. Marin, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology 142:485-501.

8. Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures in Mathematics of the Life Sciences 17:57-86.

9. Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. Journal of Molecular Evolution 39:315-329.

10. Yang, Z. 1995. On the general reversible Markov process model of nucleotide substitution: a reply to Saccone et al. Journal of Molecular Evolution 41:254-255.

11. Saccone, C., C. Lanave, G. Pesole, and G. Preparata. 1995. The reversible stationary Markov process for estimating the pattern of nucleotide substitution: a response to Ziheng Yang. Journal of Molecular Evolution 41:253.