Distances between sequences

The simplest model of nuceotide substitution is the Jukes-Cantor model. Here we assume that all bases are equally frequent, and that one base is equally likely to be substituted for another. That is, A is equally likely to be subtituted by C, G, or T. In the model, the rate at which A is replace by a non-A is u, so that the rate at which it is replaced by C, G, or T is $u/3$ for each one.

From Felsnstein 2004:

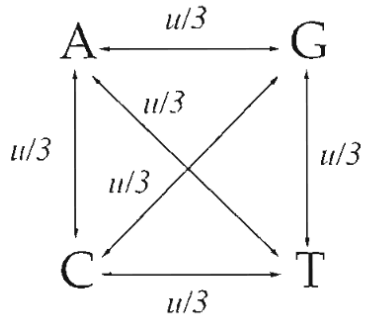


Figure 11.3: The Jukes-Cantor model of DNA change. The rate of change between all pairs of nucleotides is $u/3$ per unit time.

Another way of thinking of this is that there is a rate of $(4/3)u$ of a an event, but any of the four bases is equally likely, including A being replaced by A.

Time is treated as continuous so that the number of events along a branch is treated as a Poisson random varaible. Thus the expected number of events is

$$(4/3)ut$$

where t is the branch length.

From the Poisson distribution, we have X events having probability mass of

$$P(X = 0) = e^{-\lambda} = e^{-4ut/3}$$

Note that there are two ways for an A to remain an A: either no event occurs, or one event occurs in which A is replaced by A.

The probability of at least one substitution event is

$$P(X > 1) = 1 - P(X = 0) = 1 - e^{-4ut/3}$$

Given that there is at least one event, the probability that the last event results in a particular nucleotide is $1/4$ because the last event is equally likely to be a substitution to any of the four nucleotides. Following Felsnstein, the probability that a given taxon has a C at a particular position given that it was A at the root is

$$P(C|A, u, t) = \frac{1}{4}(1 - e^{-4ut/3})$$

Actually, the starting value of A doesn't matter. This is also the probability of observing a C regardless of the initial value.

An interesting question is what is the probability that the starting value is different from the ending value along a branch. The previous example gave the probability of a C if the starting value was an A. The same probability works for G and T, so the probability of a difference is

$$P(C|A,u,t) = \frac{3}{4}(1 - e^{-4ut/3})$$

The next slide shows a graph of this equation, which is the expected number of differences per site (which is also the probability of a difference because whether or not there is a difference is a Bernoulli random variable).

From Felsnstein 2004:

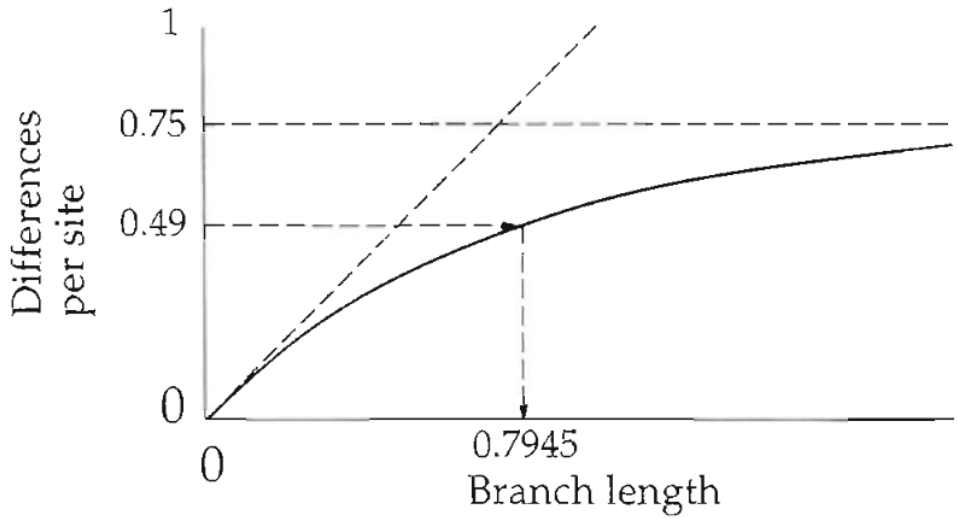


Figure 11.4: The expected difference per site between two sequences in the Jukes-Cantor model, as a function of branch length (the product of rate of change and time). The process of inferring the branch length from the fraction of sites that differ between two sequences is also shown.

In the slide, the curve asymptotically approaches 0.75 because as the time goes to infinity, the character state at that site becomes uniformly random, equally likely to be any of the four DNA letters.
This means that lengths of time of branches are not really additive.
Suppose you have taxa d and e connected by a branch, and the time to the MRCA is 0.9 units, and let $u = 0.01$. The probability that taxon d has a different value from the ancestor is

$$\frac{3}{4}(1 - e^{-4ut/3}) = \frac{3}{4}(1 - e^{-4(.01)(0.9)/3}) = 0.008946215 \approx 0.01$$

The probability e is a different from the ancestor is also approximately 0.01.

What is the probability that d and e have different DNA letters?

We can get this probability by considering the total length of the path from d to e, which 1.8 units of time. Then we have

$$\frac{3}{4}(1 - e^{-4ut/3}) = \frac{3}{4}(1 - e^{-4(.01)(1.8)/3}) = 0.01778572$$

which is less than twice the probability of a change on either of the two branches. We should expect this because it is possible that there was a change on both branches that lead to the same character state.

We can let this formula

$$D_S = \frac{3}{4}(1 - e^{-4ut/3})$$

be used as a distance between two taxa and let it represent the expected amount of change (or probability of a difference) between two taxa at a particular position in the DNA. The total distance between two taxa over an alignment will be these distances times the alignment length. This makes the often unrealistic assumption that the positions in the DNA alignment are independent. Although this seems unreasonable, it is what is normally done, and might be more reasonable for DNA that doesn't code for a protein (DNA in introns or between expressed genes in the genome). This **non-coding** DNA changes more frequently than coding DNA and is more often used for closely related species.

We can think of u and t as both being unknown parameters. But these distances only depend on u and t through their product, ut, so we can typically only estimate the product. The parameters themselves are individually not identifiable. In some cases, a value for u might be assumed from laboratory studies on mutation rates, or time might be assumed based on fossil data, so that the other parameter can be estimated.

To solve for the product ut, suppose we observe that the mean number of differences between two sequences is D. Then set

$$
\begin{aligned}
D &= \frac{3}{4}(1 - e^{-4ut/3}) \\
&\Rightarrow \frac{4}{3}D = 1 - e^{-4ut/3} \\
&\Rightarrow e^{-4ut/3} = 1 - \frac{4}{3}D \\
&\Rightarrow -4ut/3 = \log\left(1 - \frac{4}{3}D\right) \\
&\Rightarrow ut = -\frac{3}{4}\log\left(1 - \frac{4}{3}D\right)
\end{aligned}
$$

Thus

$$\widehat{ut} - \frac{3}{4}\log\left(1 - \frac{4}{3}D\right)$$

is a reasonable estimator. What kind of estimator is this?

Method of moments. (Set the theoretical mean equal to the observed sample average and solve for the parameter.) But it also works out to be the MLE of the parameter ut. As written, the estimator can actually be greater than $3/4$, so you could cap the value at $3/4$. In practice, this isn't really a probably because we don't normally work with nearly random looking sequences.

Branches of trees are often given in these units of ut, the expected number of mutations. Note that the units for u are substtitutions per unit of time, and multiplying by time, the units are the number of mutations.

The model we just used is called the Jukes-Cantor model, based on a 1969 paper. It is the simplest 4-state substitution model. Models of varying complexity have been proposed. There are many of them with different numbers of parameters. Some models are nested, so that one model is a submodel of another, in which case likelihood ratios can be used to choose between models. More generally, AIC and variants can be used to determine which substitution model best fits the data.

We'll look at a generalization called the Kimura 2-parameter model (K2P) from 1980. This model assumes one rate for DNA letters to stay within the same class: purines (A and G) and pyrimidines (C and T).

From Felsnstein 2004:

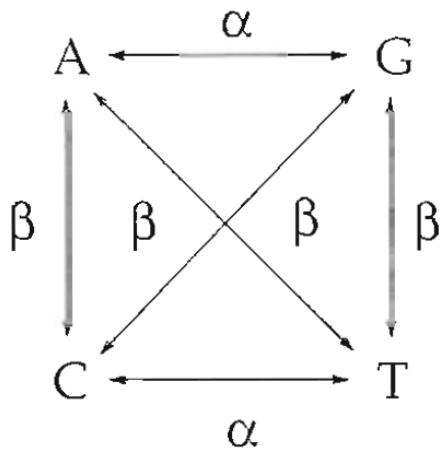


Figure 13.1: Kimura's two-parameter model. Rates differ between transitions and transversions, but are otherwise equal.

Here α and β are taking the role of u in the JC model, where α is the rate of transitions (substitutions within the same class) and transversions (substitutions between classes). For axample, A $\rightarrow$ G is a transition, and A $\rightarrow$ C is a transversion. For any given character state, there is one possible transition, and two possible transversions, so the total rate of substutions is $\alpha + 2\beta$. Biologists also use $R = \alpha/(2\beta)$, the transition/transversion ratio.

Scaling the rate to be 1, we get $\alpha + 2\beta + 1$. These three parameters then have the relationship

$$\alpha = \frac{R}{R+1}$$
$$\beta = \frac{1}{(2R+1)}]$$

So that everything can be expressed in terms of R.

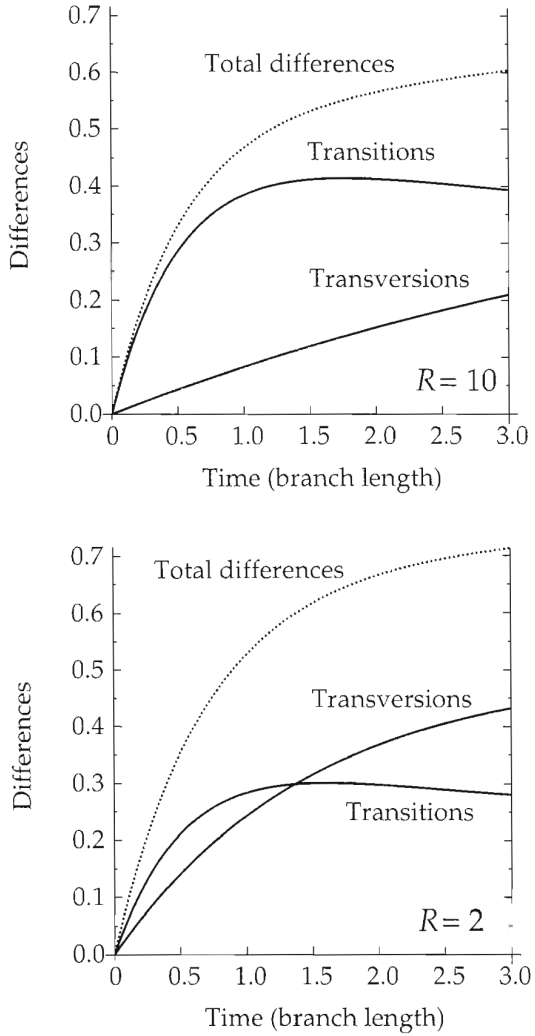
From Felsnstein 2004:

$$\begin{aligned}
\text{Prob}\,(\text{transition}\,|\,t) &= \tfrac{1}{4} - \tfrac{1}{2}\exp\left(-\tfrac{2R+1}{R+1}\,t\right) + \tfrac{1}{4}\exp\left(-\tfrac{2}{R+1}\,t\right) \\
\text{Prob}\,(\text{transversion}\,|\,t) &= \tfrac{1}{2} - \tfrac{1}{2}\exp\left(-\tfrac{2}{R-1}\,t\right)
\end{aligned} \tag{13.2}$$

With the Jukes-Cantor distance, we could compute the distance by expressing the fraction of difference between sequences in terms of the distance, and then solving that equation for the distance. With Kimura's two-parameter distance, a similar procedure can be followed. In effect, we have two observations, the fraction P of transition differences between the two sequences, and the fraction of transversion differences (Q). Solution of the two equations 13.2 yields:

$$\begin{aligned}
\widehat{t} &= -\tfrac{1}{4}\ln\left[(1-2Q)(1-2P-Q)^2\right] \\
\widehat{R} &= \frac{-\ln(1-2P-Q)}{-\ln(1-2Q)} - \tfrac{1}{2}
\end{aligned} \tag{13.3}$$

From Felsnstein 2004:



From Felsnstein 2004:

(there are two of the latter). If there are n sites in all, n_1 of which have transition differences, and n_2 transversion differences, the product of terms will be:

$$L = \text{Prob}(\text{data} \mid t. R) = \left(\frac{1}{4}\right)^{n} (1 - P - Q)^{n - n_1 - n_2} P^{n_1} \left(\frac{1}{2}Q\right)^{n_2} \qquad (13.4)$$