Massachusetts Institute of Technology
6.877 Computational Evolutionary Biology, Fall, 2005
Notes for November 7: Molecular evolution

1. Rates of amino acid replacement

The initial motivation for the neutral theory came from observations on the rate of amino acid replacements in proteins. When extrapolated to the entire genome, the inferred rate of evolution was several nucleotide substitutions per year. This rate was regarded as much to high to result from natural selection, because the intensity of selection must be limited by the total amount of differential survival and reproduction that occurs in the organism. Direct DNA sequencing later revealed that rates of nucleotide substitution vary according to the function or presumed absence of function of the nucleotides.

The first 18 amino acids in the amino terminal end of the human and mouse gamma interferon proteins are a signal peptide that is used in the secretion of molecules. The sequences are:

```
Hum: Met Lys Try Thr Ser Tyr Ile Leu Ala Phe Gln Leu Cys Ile Val Leu Gly Ser
Mou: Met Asn Ala Thr His Tyr Cys Leu Ala Leu Gln Leu Phe Leu Met Ala Val Ser
```

We could simply count the number of sites that differ in the two sequences: there are 10/18 that differ, or 0.56.

To interpret this, let us assume that amino acid replacements occur at the rate λ per unit time. Consider two independently evolving sequences, initially identical, which at time t are found to differ in proportion D_t of their amino acids. After the next time interval, the proportion of differences is just $D_{t+1} = (1-D_t)(2\lambda)+D_t$, so we add to the already existing proportion of differences those that might have arisen in the current time interval. The factor of 2 is present because the total time for evolution is $2t$ time units (t units in each lineage). The equation ignores the unlikely possibility of an amino acid replacement making two previously different sites identical.

If λ is the rate of amino acid replacement per unit time, then the probability that a particular site remains unsubstituted for t consecutive intervals along each of two independent lineages is $(1-\lambda)^{2t}$ which is approximately $e^{-2\lambda t}$ if λt is not too large. So, the probability of 1 or more replacements is just 1 minus this quantity, or about $(1-e^{-2\lambda t})$.

Since λ is the rate of amino acid replacement per unit time, the expected proportion of differences between two sequences at any time t is $K=2\lambda t$. Substituting and rearranging gives the estimate of K from the observed differences $K = -\ln(1-D_t)$. This quantity is used in preference to D because it takes multiple substitutions into account, especially over a long period of time.

Rates of amino acid substitution vary over a 500-fold range in different proteins. The rate of amino acid replacement in gamma interferon is one of the largest rates known.

Among the slowest rates is that of histone H4, for which the substitution rate λ is 0.01 x 10^{-9} per year. The average rate for many proteins is very close to the rate found in hemoglobin, which is about 1 x 10^{-9} amino acid replacement per amino acid site per year.

So far we have done our computations on polymorphism sequences based on just the observed counts, all on the assumption that we were looking at individuals from the same species. However, when we look across species, or even within species, if the time since common divergence is long enough, there is always the possibility that the same nucleotide site position could have changed more than once. For example, position 1, say, could first be an A; then be substituted with a T; and then once more back to A. In this case, we would not 'see' any change from the ancestral state even though a change had occurred. We call this a 'multiple hit'. (So in a way this relaxes the constraint of the infinite sites model.) In general then, the number of nucleotide substitutions that we observe, and that we have been calling mutations, are always an undercount of the true number. There are many models of nucleotide substitutions that have been proposed to provide a 'correction factor' for this possibility.

Nucleotide sequences are analyzed in the same manner as amino acid sequences, but we have to correct for cases where a substitution makes two previously different nucleotide sites identical. This is significant because an expected 1/3 of random substitutions will make two previously different nucleotides identical, while the chance is only 1/19 for amino acids.

Many models of nucleotide substitution have been proposed, which differ mostly in the assumptions about rates of mutations between pairs of nucleotides. The simplest model for nucleotide substitution assumes that mutation occurs at a constant rate, and each nucleotide is equally likely to mutate to any other, the Jukes-Cantor model. If α is the rate of mutating from one nucleotide to another, then in any time interval, A mutates to C with probability α, A mutates to T with probability α, and A mutates to G with probability α. The probability that it does not mutate in this time interval is therefore $(1-3\alpha)$, The probability that a particular site is A at time $t+1$ is the probability of having been A at time t and not mutating, plus the probability of being at any other nucleotide an mutating to A: $P_{A(t+1)} = (1-3\alpha)P_{A(t)} + \alpha(1-P_{A(t)})$. From this we can obtain a simple differential equation:

$$\frac{dP_{A(t)}}{dt} = -4\alpha P_{A(t)} + \alpha$$

which as we know show has the solution:

$$P_{A(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

To solve this, we can use an integrating factor, multiplying both sides by $f(t)$:

$$f(t)4\alpha P_{A(t)} + f(t)\frac{dP_{A(t)}}{dt} = \alpha f(t)$$

we need $f'(t) = f(t)4\alpha$ so that we can use integration by parts

so if $f'(t) = f(t)4\alpha,$ then $f(t) = e^{4\alpha t}$

Substitution this value for *f*, we get:

$$e^{4\alpha t}4\alpha P_{A(t)} + e^{4\alpha t}\frac{dP_{A(t)}}{dt} = \alpha e^{4\alpha t}$$

$$\frac{d}{dt}\left(e^{4\alpha t}P_{A(t)}\right) = \alpha e^{4\alpha t}$$

$$\int \frac{d}{dt}\left(e^{4\alpha t}P_{A(t)}\right) = \int \alpha e^{4\alpha t} \quad = \alpha e^{4\alpha t} + C$$

$$e^{4\alpha t}P_{A(t)} = \alpha e^{4\alpha t} + C$$

To find the value of *C*, we use the fact that at time t=0, $P_A(0)$=1 (we start at A). This gives us:

$$1 = \frac{1}{4} + Ce^{-4\alpha 0} \text{ or } C = \frac{3}{4}$$

If we observe two sequences that have been separated for time *t*, then the probability that they continue to carry the same nucleotide at a particular site is:

$$P_{AA} = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}$$

The proportion of sites that differ between the two sequences is just 1 minus this, or D=(1–P_{AA}). Therefore,

$$D == \frac{3}{4}\left(1 - e^{-8\alpha t}\right)$$

In our previous symbols, λ is the rate of mutation to a nucleotide different from the current nucleotide, so relating this to α, we have λ=3α. This implies that K=2λt=2(3α)=6αt. Taking logs of both sides of the equation above, we get:

$$8\alpha t = -\ln\left(1 - \frac{4}{3}D\right)$$

and since K=3/4(8αt),

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right)$$

This is the Jukes-Cantor correction factor. If one lets *t* go to infinity, the divergence will approach ¾ , which makes sense: after enough time, the common ancestry of the two sequences has been erased, and ¼ of all sites will match by chance.

Example 1.

If 25% of observed sites are different, D=0.25, then plugging this into the formula, the actual number of substitutions is K=0.3404, higher, as expected.

Felsenstein's intuitive picture.

To construct a more general picture, we can posit a 4x4 transition probability matrix, whose entries give the probability that a particular site will change to another site in one time-step. Let u be the mutation rate per unit per site; P_{ij}= the probability that base i changes to base j in one generation. If t is large and u is small, we can approximate the process as a Poisson, on a continuous scale, with a constant risk of mutation. Note that if we had a type of mutation that changed a base to any one of the four bases chosen at random, with equal probability of the four outcomes, it would almost be like Jukes-Cantor, except that ¼ of the time it would make no change at all (it would do nothing, i.e., no event would occur at all). Now we imagine that we alter this new model by increasing the mutation rate to $4/3u$. In this (rescaled) model, the mutations that change the bases occur at rate u, continuously in time (this includes the events that change, e.g., a base A to the same base A).

Now this is just a Poisson model with mean rate of change $4/3u$, so we can write the probability that k mutations occur, given that time t has passed, as follows:

$$\Pr[k \mid t] = e^{-\frac{4}{3}ut} \frac{\left(\frac{4}{3}u\right)^k}{k!}$$

Now we can easily compute the probability that we wind up with base j having started with a base i. The probability that zero base-changing events occur at all during time t is just the 0^{th} term of the Poisson, which is just $exp(-4/3ut)$ (The $(4/3u)^j$ term drops out when j=0, as usual.). Therefore, the probability that there is at least one event that picks a base ¼ at random is 1–this probability, or $(1-exp(-4/3ut))$. ¼ of the time this change will be to some particular base. Putting this all together, we have:

$$P_{ij} = \frac{1}{4}(1 - e^{-\frac{4}{3}ut}) \quad (\text{for } i \neq j)$$

$$P_{ij} = \frac{1}{4}(1 - e^{-\frac{4}{3}ut}) + e^{-\frac{4}{3}ut} \quad (\text{for } i = j)$$

because in the second case we must add the probability that no base-changing event has happened at all, leaving us at base i. Thus, if we add up the expected proportion of sites that two sequences will differ after time t, it is just 3 times the first case above, giving us the expected fraction of sites that differing:

$$D == \frac{3}{4}\left(1 - e^{-\frac{4}{3}ut}\right)$$

Note that the re-parameterization via the ut term is meant to replace the parameter K in the earlier derivation: the value of t here is actually the time $2t$ (total divergence time since common ancestor) given earlier.

The Kimura 2-parameter model.

The Jukes-Cantor model suffers from several biological inaccuracies. For one thing, all transitions are assumed equal. However, it is well known that the rate of mutations between the two chemically similar purine bases (Adenine and Guanine, A and G) or between the two chemically similar pyrimidine bases (Cytosine and Thymine, C and T), is much higher (generally more than double) the rate of mutations between a purine and a pyrimidine (e.g., from an A to a T). This is particularly true, e.g., in mammals. The purine-purine or pyrimidine-pyrimidine mutations are called transitions and the purine-pyrimidine mutations, transversions. If there were only one mutation parameter, as in the JC model, we would expect a 2:1 ration of transversions to transitions. Estimates from mammalian sequences, however, show that 57% of changes are actually transitions.

There are two possible transition pairs (A-G and C-T) and four possible transversion pairs (A-T, A-C, G-T, G-C), discounting self-loops. So, in all there are 4 possible transitions, 8 transversions, and 4 self-loops. This suggests that one ought to move from a one parameter, single rate model to at least a 2 parameter rate model, one for the transition mutation rate (often denoted t_s) and one parameter for the transversion mutation rate (often denoted t_v). This is what Kimura suggested; this model is known as K2P. In the formulations of this model, we will often see the difference between the transition mutation rate and the transversion mutation rate expressed as a ratio, R= transition/transversion rates, or $R=t_s/t_v$.

We can set up a differential equation as before to solve for the K2P model and its correction factor. If the time divergence is not too great, this will be close to Jukes-Cantor. We can also solve this model using general methods for continuous time Markov chains, as shown later for a slightly different model. Or, one can do the same sort of reparameterizing 'trick' as with the Jukes-Cantor model, which we also defer to the next section. For now, we simply give the solution. If P denotes the total frequency of transitions and Q the total probability of transversions then the K2P says that:

$$P = (1/4)(1 - 2e^{-4(\alpha+\beta)t} + e^{-\beta t})$$

$$Q = (1/2)(1 - e^{-8\beta t})$$

$$K = 2\lambda t = 2\alpha t + 4\beta t =$$

$$-(1/2)\ln(1 - 2P - Q) - (1/4)(1 - 2Q)$$

For the Kimura model, we can show that the equilibrium frequency for all bases is ¼, irrespective of their initial frequency – this follows from the symmetry of the model, as in Jukes-Cantor. However, turning this around, also like Jukes-Cantor, the formulas are applicable irrespective of the initial base frequency, so the model is applicable to a wider range of cases than many other models that are more complex. Further, the more complex models require one to estimate more parameters. We'll probe this further a bit later.

Example. You are given two sequences (50 bp each) of the homologous pseudogene in two species of yeast. (Pseudogene: non-transcribed gene that has 'decayed' and is subject to purely neutral evolution.) The alignment is shown below:

```
Species 1: CCTCGACGGCTTAGATCTGATCTGACCTAATGCTGCAATCGTTACAAAGT
Species 2: CCTCCACGAGTAAGAGTTGATCCGACTTAGTCCTGCGATCGTTAGATAAT
```

Note that Jukes-Cantor isn't exactly applicable here because there are 7 transitions and 7 transversions, 14 substitutions out of 50 base pairs. However, the Jukes-Cantor model assumes that there are twice as many transversions as transitions. To apply the Kimura correction, if we knew, say, that the two species diverged 10 million years ago, and there are 50 generations per year, the divergence time is 2 x 500 million or 10^9, but we'd also have to get estimates of the transition/transversion ratio. In this data, it looks to be 1.

The Jukes-Cantor correction would proceed as follows: There are 14 substitutions in 50 bp. D= 14/50 = 0.28. K= (3/4)*ln(1-(4/3)*D)=0.35. You should figure out what the difference is for the Kimura model, given R=1. Finally, notice that if we have these figures, we can determine the mutation rate (assuming that it is equal for the two species):

$K = 2t$ (years)*μ(mutations per bp per year)
μ = 0.35/20 Myr = 1.75×10^{-8} mutations per bp per year

Forthcoming: the matrix method of solution for estimating substitution corrections.