Lecture 10 – Models of DNA Sequence Evolution

I. Introduction. So, given the central role that model-based methods play in phylogeny estimation, we need to understand models of sequence evolution.

These are important in two contexts. First, they're used to correct for multiple substitutions in calculating pairwise genetic distances (substitutions per site between a pair of sequences) for the minimum evolution criterion.

Second, they're central to methods that use likelihood as an optimality criterion because it's from these models that we derive the transformation probabilities. Remember from earlier, the likelihood of a particular reconstruction is as follows:

$$P(R_r \mid \tau) = \pi_{\mathrm{m}} \times P_{m,k}(v_{3,1}) \times P_{k,\mathrm{A}}(v_{1,w}) \times P_{k,\mathrm{G}}(v_{1,x}) \times P_{m,l}(v_{3,2}) \times P_{l,\mathrm{C}}(v_{2,y}) \times P_{l,\mathrm{C}}(v_{2,z})$$

and we can summarize them into the SSL as follows:

$$L^{i}_{(\tau)} = \sum_{m} \pi_m \times \left(\sum_{k} P_{m,k}(v_{3,1}) P_{k,G}(v_{1,w}) P_{k,A}(v_{1,x}) \right) \times \left(\sum_{l} P_{m,l}(v_{3,2}) P_{l,C}(v_{2,y}) P_{l,C}(v_{2,z}) \right)$$

It's the $P_{i,j}$'s, the probability of a substitution from nucleotide i to j, that we need a substitution model to calculate.

Typically, a Markov process is used to model sequence evolution. In particular, we'll assume a Poisson process to model substitutions, were time between events in exponentially distributed, with rate λ.

II. Jukes-Cantor.

The first model that was developed to account for multiple substitutions is the Jukes-Cantor Model. (This derivation is from Li, W.-H. 1997. *Molecular Evolution*, Sinauer)

Under this model, the probability of a site remaining constant is:

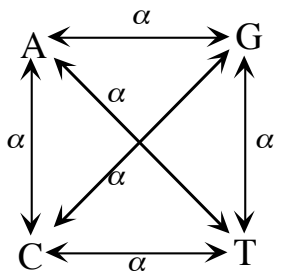
$$p_{ii(t)} = \tfrac{1}{4} + \tfrac{3}{4}\, e^{-4\alpha t}$$

The probability of a site changing is given by:

$$p_{ij(t)} = \tfrac{1}{4} - \tfrac{1}{4}\, e^{-4\alpha t}$$

We'll spend a few minutes understanding this, as it illustrates the logic of models of sequence evolution.

Let's start as follows:



α is the rate at which any nucleotide changes to any other per unit time (analogous to λ).

Given that the state at the site is A at t_0, we start by estimating the probability of an A at that site at t_1:

$$p_{A(0)} = 1 \text{ (Given)}$$
$$p_{A(1)} = 1 - 3\alpha,$$

because there are three ways for A to change, each of which occurs at rate α.

Now, what's the probability of this site having an A at t_2?

There are two ways for the site to have an A at t_2:

1 – It still hasn't changed.
2 – It has changed to something else and back again (i.e., *it experienced a multiple hit*).

Therefore,

$$p_{A(2)} = (1 - 3\alpha)\, p_{A(1)} + \alpha\,(1 - p_{A(1)})$$

$(1 - 3\alpha)\, p_{A(1)}$ = probability of no change at the site during time 2, $(1 - 3\alpha)$, times the probability of the site having A at time 1, $p_{A(1)}$.

$\alpha\,(1 - p_{A(1)})$ = probability of a change to A, α, times the probability that the site is not *A* at time t_1, $(1 - p_{A(1)})$.

So we essentially have a recursion equation:

$$p_{A(t+1)} = (1 - 3\alpha)\, p_{A(t)} + \alpha\,(1 - p_{A(t)}) = p_{A(t)} - 3\alpha p_{A(t)} + \alpha\,(1 - p_{A(t)})$$

Next, we can calculate the change in $p_{A(t)}$ across a single unit of time.

$$p_{A(t+1)} - p_{A(t)} = -3\alpha p_{A(t)} + \alpha\,(1 - p_{A(t)}),$$

$$\text{so } \Delta p_{A(t)} = -3\alpha p_{A(t)} + \alpha (1 - p_{A(t)}) = -3\alpha p_{A(t)} + \alpha - \alpha p_{A(t)}$$

$$\Delta p_{A(t)} = -4\alpha p_{A(t)} + \alpha$$

We can express this as a continuous-time process with the following:

$$\frac{dp_{A(t)}}{dt} = -4\alpha p_{A(t)} + \alpha$$

or,

$$p_{A(t)} = 1/4 + (p_{A(0)} - 1/4)\, e^{-4\alpha t}$$

We have a probability that a site has a particular nucleotide after time *t,* given in terms of its initial state. Remember that under JC, all transformations occur at the same rate, so we can generalize this as follows.

If $i = j$, $p_{A(0)} = 1$. Therefore,

$$p_{ii(t)} = 1/4 + 3/4\, e^{-4\alpha t}$$

If $i \neq j$, $p_{A(0)} = 0$ and

$$p_{ij(t)} = 1/4 - 1/4\, e^{-4\alpha t}$$

Now α is an instantaneous rate, so we've modeled branch length (rate times time) explicitly in our expectations.

Note that in his text, Felsenstein expresses α as $\mu/3$.

The JC model makes several assumptions.

1) All substitutions are equally likely. All are treated equally, and we have a single substitution type (nst = 1).

2) Base frequencies are assumed to be equal. That is, each of the four nucleotides occurs at 25% of sites (ba = eq).

3) Each site has the same probability of experiencing a substitution as any other. That is, all nucleotide sites are evolving the same rate (ra = eq).

4) The process is constant through time.

5) Sites are independent of each other.

6) Substitution is a Markov process. The probability of a particular change at a site is not dependent on the prior history of that site

Some of these assumptions are quite bad, and models have been developed to relax many of them.

The first two can be relaxed by the elaboration of the Q-matrix.

III. Substitution types and base frequencies.

We can think about the JC model in terms of a matrix of substitution rates. This matrix is called the **Q** matrix, and essentially represents the instantaneous rates for each substitution type. The diagonals are chosen so the rows sum to 0.

For J.C., $\mathbf{Q} =$

$$\begin{matrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha. \end{matrix}$$

A. The GTR model. We can generalize the **Q** matrix as follows:

$$\mathbf{Q} = \begin{matrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G - e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{matrix}$$

where, μ = the average instantaneous substitution rate,
$a, b, c, \ldots, l$ are relative rate parameters (one of them is set to 1).
and π_i's are the frequencies of the base that is being substituted *to*.

Note that this is not symmetric, and therefore, the full model is non-reversible.

In almost all cases, we use reversible models, and this is accomplished simply by setting the following:

$$a = g, b = h, c = i, d = j, e = k, \& f = l.$$

Thus, for the most general time-reversible model (conveniently called the GTR model), we have the following matrix of instantaneous substitution probabilities:

$$\mathbf{Q} = \begin{matrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{matrix}$$

Although this is not actually a symmetric matrix, it can be decomposed into two symmetric matrices, usually called **R** & **Π**.

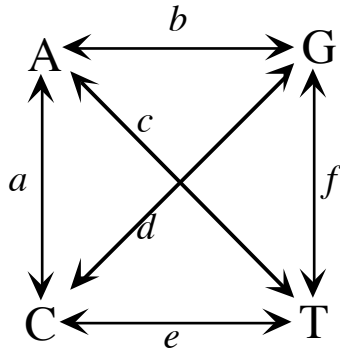
$$\mathbf{R} = \begin{matrix} --- & \mu a & \mu b & \mu c \\ \mu a & --- & \mu d & \mu e \\ \mu b & \mu d & --- & \mu f \\ \mu c & \mu e & \mu f & --- \end{matrix}$$

and

$$\mathbf{\Pi} = \begin{matrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{matrix}$$

The R-matrix is something we'll spend a fair amount of time with, as it's something we want to tailor to analyses of individual data sets. The rate parameters of the R-matrix are values that we input in likelihood analyses (as are the base frequencies), or we can optimize them.

So, we can think of the GTR model in visual terms as follows:



where $a - f$ represent the relative-rate parameters, which are multiplied by the mean substitution rate.

B. Common Simplifications.

As we've mentioned, for most genes, there is actually a much higher probability of transition type substitutions than transversion substitutions.

The **Kimura two-parameter (K2P) model** was the first to deal with this observation.

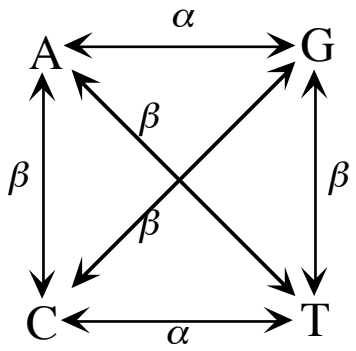
So we set $b = e = \kappa$ (for transitions), and $a = c = d = f = 1$ (for transversions). There are two substitution types.

Like JC, the K2P model assumes equal base frequencies, so all π_i's = ¼

for K2P: $\mathbf{Q} =$

$$\begin{pmatrix} -(\mu)(\kappa+2)/4 & \mu/4 & \mu\kappa/4 & \mu/4 \\ \mu/4 & -(\mu)(\kappa+2)/4 & \mu/4 & \mu\kappa/4 \\ \mu\kappa/4 & \mu/4 & -(\mu)(\kappa+2)/4 & \mu/4 \\ \mu/4 & \mu\kappa/4 & \mu/4 & -(\mu)(\kappa+2)/4 \end{pmatrix}$$

where $\alpha = \mu\kappa/4$ and $\beta = \mu/4$. Thus, $\kappa = \alpha / \beta$ where,



and the Q matrix for the K2P model can be represented as such.

$$\text{for K2P: } \mathbf{Q} = \begin{matrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{matrix}$$

So if $\kappa = 4$, we expect there to be twice as many transitions as transversions.

Remember that here we set all base frequencies equal to 0.25. It's commonly the case that actual base frequencies deviate from this substantially.

The **Hasegawa-Kishino-Yano (HKY) Model** extends the K2P to unequal base frequencies.

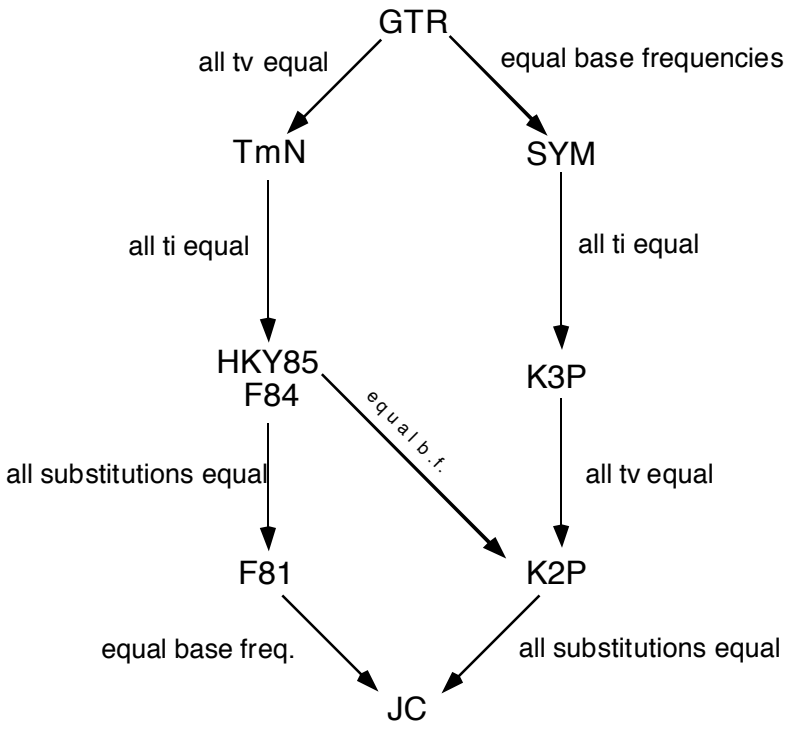
$$\text{for HKY: } \mathbf{Q} = \begin{matrix} -\mu(\kappa\pi_G + \pi_Y) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\kappa\pi_T + \pi_R) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C + \pi_R) \end{matrix}$$

where $\alpha = \mu\kappa$, $\beta = \mu$, $\pi_R = \pi_A + \pi_G$, and $\pi_Y = \pi_C + \pi_T$.

Another method for the same conditions was given by Felsenstein in 1984 and is called the F84 model.

There are lots of other named models that restrict the cells of the **Q** matrix for the GTR model.

Figure 11 from Swofford et al. (1996).



This really represents a family of models (called the GTR family) because there are many ways we can restrict the cells of this matrix to give us the familiar models. In addition, there are lots of ways that we could restrict the relative rate parameters of the GTR matrix to employ models that haven't been described or named.

So, each of the simpler models is a special case of the GTR model. We'll actually use this fact later.

We can deal with this by assigning rate classes that restrict the R-matrix. rclass = (a b a a c a), allows the two transitions to have unique relative rates, but there's a single tv rate.

IV. Calculating transformation probabilities.

So the Q & R matrices we've been discussing define the instantaneous rates of substitutions from one nucleotide to another.

We still have to convert the rates to probabilities before we can use the models either to generate distances that are corrected from multiple hits or for calculating the likelihood of a particular tree for a site.

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

For the simple models, this is easy enough to compute.

I've already shown you the probabilities associated with the **JC model:**

$$P_{ij(t)} = \begin{cases} 1/4 + 3/4\, e^{-4\alpha t} & \text{for } i = j \\ 1/4 - 1/4\, e^{-4\alpha t} & \text{for } i \neq j \end{cases}$$

There are similar formulae for some of the other simpler models, for example K2P:

$$P_{ij(t)} = \begin{cases} 1/4 + 1/4\, e^{-\mu t} + 1/2\, e^{-\mu t((\kappa+1)/2)} & (\text{for } i = j) \\ 1/4 + 1/4\, e^{-\mu t} - 1/2\, e^{-\mu t((\kappa+1)/2)} & (\text{for } i \neq j, \text{transition}) \\ 1/4 - 1/4\, e^{-\mu t} & (\text{for } i \neq j, \text{transversion}) \end{cases}$$

The formulae for the transformation probabilities of the HKY model are given in the Swofford et al. (1996) chapter of *Molecular Systematics* (Chapter 11). For the more complicated of the GTR family of models, the eigenvalues and eigenvectors of the **Q** matrix must be calculated by numerical evaluation.

Next, we'll deal with incorporating rate variation among sites to our models and we'll discuss ways that have been developed to select a model.