The Markov model of sequence evolution **Tuesday, September 24th**

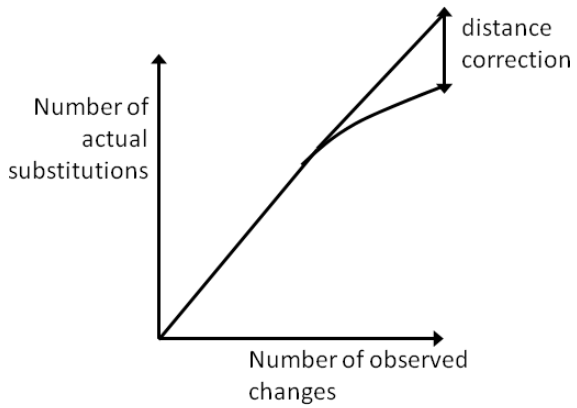
Dannie Durand

The process of substitution at a single site in a nucleotide sequence can be modeled as a Markov chain where each state represents a single nucleotide and the transition probability, P_{jk}, is the probability of replacing nucleotide j with nucleotide k. Similarily, Markov chains can be constructed to model the evolution of amino acid sequences.

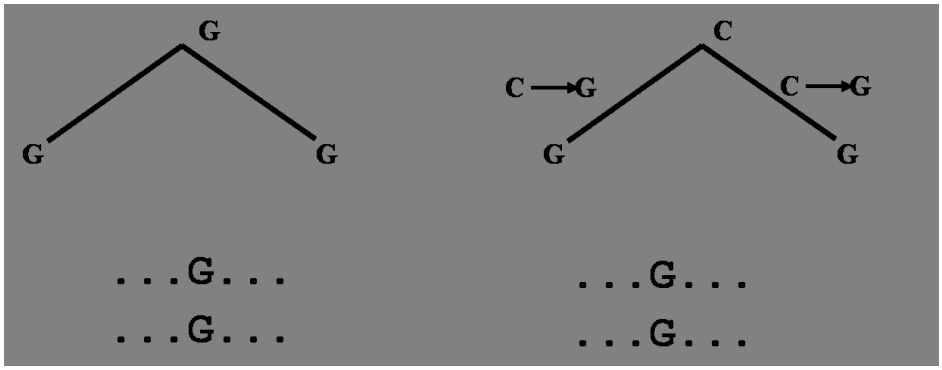
Markov models of sequence substitution are used to answer a wide range of questions that arise in molecular evolution:

- Correcting for multiple substitutions
- Simulating sequence evolution
- Estimating rates of evolution
- Deriving substitution scoring matrices
- Estimating the likelihood of observing a pair of aligned nucleotides, given a phylogenetic model.

In today's lecture, we discussed estimating the number of substitutions that occurred at a given site. In molecular phylogenetics, distances between taxa are typically calculated from a multiple alignment. Multiple substitutions at the same site are a major source of inaccuracy in such distance estimates. If only a few changes have occurred, then the observed number of mismatches may, in fact, be the actual number of substitutions. However, as the divergence increases, so does the probability of two or more substitutions at the same site. In this case, the number of observed changes will underestimate the actual distance as shown below:



Suppose that in a pairwise alignment of two sequences, σ and τ, we wish to estimate the number of substitutions that actually occurred over Δt, the time interval that elapsed since they diverged from a common ancestor. For example, if there is a G at the same position in both sequences, it could be because the ancestral state was also G and no change occured (left hand figure) or because parallel changes occured in both sequences (right hand figure).



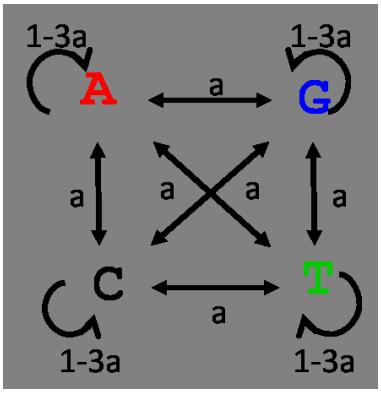
In today's lecture, we described how to use the Jukes Cantor model to estimate this quantity. If we assume a constant rate of substitution, λ, in both lineages then the number of substitutions per site is $2\lambda\Delta t$. Both λ and Δt are unknown. However, we do know the fraction of positions that are not identical in the present-day alignment. We will use the observed number of mismatches and a Markov model and sequence substitution, to estimate $2\lambda\Delta t$.

The Jukes-Cantor model

We define a Markov model of substitution in a single time step. The simplest such model for DNA is the Jukes-Cantor model[1], which assumes that all substitutions ($A \to C, A \to G, A \to T, C \to A...$) are equally likely and occur at a rate α. The consequence of this assumption that the overall rate of substitution is $\lambda = 3\alpha$.

The transition probability matrix for this Markov model is:

[1] Jukes and Cantor, Evolution of protein molecules. In H. N. Munro, (ed.) *Mammalian Protein Metabolism*, 21-123, Adademic Press, NY, 1969.



$$\begin{bmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{bmatrix}$$

The stationary distribution of this Markov chain is $\varphi = (0.25, 0.25, 0.25, 0.25)$.

Using this Markov chain we derive an expression describing how changes accumulate at site i over a period of time Δt. The probability of observing, for example, an A at site i after one time step has elapsed is given by

$$\varphi_A^{(t+1)} = (1-3\alpha)\varphi_A^{(t)} + \alpha\varphi_G^{(t)} + \alpha\varphi_G^{(t)} + \alpha\varphi_T^{(t)}$$

where $\varphi_j^{(t)}$ is the probability of being in state E_j at time t. This reduces to

$$\varphi_A^{(t+1)} = (1-3\alpha)\varphi_A^{(t)} + \alpha\left[1-\varphi_A^{(t)}\right].$$

Here, the first term gives the probability of observing A at time $t+1$ if the residue at site i at time t was an A. The second term is the probability of observing A if the residue at time t was not an A. Since the model is symmetric, this equation applies equally well to C, G or T. We can therefore rewrite the equation using the parameter y, where $y \in A, C, G, T$, yielding

$$\varphi_y^{(t+1)} = (1-4\alpha)\varphi_y^{(t)} + \alpha.$$

After subtracting $\varphi_y^{(t)}$ from both sides, and some algebraic manipulation we obtain

$$\varphi_y^{(t+1)} - \varphi_y^{(t)} = \alpha\left(1-4\varphi_y^{(t)}\right).$$

Applying a continuous time approximation allows us to express this as a differential equation

$$\frac{d\varphi_y^{(t)}}{dt} = \alpha\left(1 - 4\varphi_y^{(t)}\right)$$

with solution

$$\varphi_y^{(t)} = \frac{1}{4} + \left(\varphi_y^{(0)} - \frac{1}{4}\right) e^{-4\alpha t}.$$

We now have an expression for the probability of observing nucleotide y at site i after an arbitrarily long elapsed time, Δt. We have two cases. The probability that the present-day residue is the same as the ancestral nucleotide ($\varphi_{y=x}^{(0)} = 1$) after time Δt is

$$p_{xx}(\Delta t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha\Delta t}. \tag{1}$$

The probability that the present-day nucleotide differs from the ancestral residue ($\varphi_y^{(0)} = 0$) is

$$p_{yx}(\Delta t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha\Delta t}. \tag{2}$$

The next step is to estimate the expected number of observable differences (mismatches) between the two sequences. First, we derive an expression for the probability of observing a match, for example, for observing two adenines aligned at site i. Given two sequences evolving independently from a common ancestral sequence with an unknown nucleotide x at site i, the probability that both sequences will have an A at site i is

$$P_M = \left[p_{AA}^{(\Delta t)}\right]^2 + \left[p_{TA}^{(\Delta t)}\right]^2 + \left[p_{CA}^{(\Delta t)}\right]^2 + \left[p_{GA}^{(\Delta t)}\right]^2,$$

where Δt is the elapsed time since their divergence. Replacing the first term with equation (1) and the remaining terms with equation (2), this reduces to

$$P_M = \left[\frac{1}{4} + \frac{3}{4} e^{-4\alpha\Delta t}\right]^2 + 3\left[\frac{1}{4} + \frac{1}{4} e^{-4\alpha\Delta t}\right]^2.$$

The first term gives the probability of observing A's in both sequences if the ancestral nucleotide was also A. The second term represents the case where the ancestral nucleotide was not an A. By expanding the squared quantities and combining terms, we obtain

$$P_M = \frac{1}{4} + \frac{3}{4} e^{-8\alpha\Delta t}. \tag{3}$$

Note that since the Jukes Cantor model is symmetric, equation (3) in fact gives the probability of observing the same nucleotide x in both sequences, where x may be any nucleotide. P_m, the probability of observing a mismatch at site i, is simply 1 minus the probability of a match or

$$\begin{aligned} P_m &= 1 - P_M \\ &= \frac{3}{4}(1 - e^{-8\alpha\Delta t}). \end{aligned}$$

Recall that our ultimate goal is to estimate the expected number of substitutions that occurred at site i since the sequences diverged. This quantity is $E[sub] = 2\lambda\Delta t = 6\alpha\Delta t$. We solve the above equation to obtain an expression for $E[sub]$ in terms of P_m:

$$\alpha\Delta t = -\frac{1}{8}\, ln\, (1 - \frac{4}{3}P_m).$$

Multiplying both sides of the equation by 6, we obtain an expression for the expected frequency of substitutions per site, in terms of the number of sites with an observable difference:

$$E[sub] = -\frac{3}{4} ln\, (1 - \frac{4}{3}P_m).$$

If we estimate the expected number of observable differences by the number of differences actually observed, m/n, then

$$E[sub] = -\frac{3}{4} ln\, (1 - \frac{4}{3}\,\frac{m}{n}).$$

So, for example, if we observe mismatches at 100 sites in a nucleotide sequence of length 1,000, then the Jukes-Cantor model predicts that the actual number of substitutions per site is 0.107 or 107 substitutions.

More complex models of nucleotide substitution

The rate of each possible substitution, α is an explicit parameter of the Jukes-Cantor model. The frequencies of A's, G's, C's and T's are implicitly specified by the model, since this is determined by the stationary distribution.

Nucleotide substitution models can be made more realistic in two directions. First, the assumption that all sustitutions occur at the same can be relaxed. For example, the Kimura 2 Parameter model

assumes that transitions and transversions occur at different rates:

$$\begin{bmatrix} & A & C & G & T \\ A & 1-\alpha-2\beta & \beta & \alpha & \beta \\ C & \beta & 1-\alpha-2\beta & \beta & \alpha \\ G & \alpha & \beta & 1-\alpha-2\beta & \beta \\ T & \beta & \alpha & \beta & 1-\alpha-2\beta \end{bmatrix}$$

Second, the specification of the rates can be adjusted to yield a non-uniform stationary distribution. The Felsenstein (1981) model, like the Jukes-Cantor model, assumes that all substitutions are equally likely, but can model an stationary distribution, $\varphi = (\varphi_A, \varphi_C, \varphi_G, \varphi_T)$:

$$\begin{bmatrix} & A & C & G & T \\ A & 1-\alpha\cdot(\varphi_C+\varphi_G+\varphi_T) & \varphi_C\cdot\alpha & \varphi_G\cdot\alpha & \varphi_T\cdot\alpha \\ C & \varphi_A\cdot\alpha & 1-\alpha\cdot(\varphi_A+\varphi_G+\varphi_T) & \varphi_G\cdot\alpha & \varphi_T\cdot\alpha \\ G & \varphi_A\cdot\alpha & \varphi_C\cdot\alpha & 1-\alpha\cdot(\varphi_A+\varphi_C+\varphi_T) & \varphi_T\cdot\alpha \\ T & \varphi_A\cdot\alpha & \varphi_C\cdot\alpha & \varphi_G\cdot\alpha & 1-\alpha\cdot(\varphi_A+\varphi_C+\varphi_G) \end{bmatrix}$$

The Hasegawa, Kishino, Yano (HKY) model, which combines both innovations, allows different rates for transitions and transversions and an arbitrary stationary distribution, $\varphi = (\varphi_A, \varphi_C, \varphi_G, \varphi_T)$. The most general in this family of models, the General Time Reversible (GTR) model, allows a different rate for each of the six possible substitutions and an arbitrary stationary distribution.

In deciding which model to use for a particular data set, we face the usual tradeoff: more general models can give a better fit, but require more data to infer more parameter values and have a greater danger of overfitting. In addition, these models do not allow for changes in rate or in GC-content over time.