## Felsenstein & the birth of statistical phylogenetics

I. Why do we need statistical phylogenetics? One of the most well-documented phenomena in molecular evolution is that of *saturation*. That is, the relationship between time and the raw number of differences between two sequences (*uncorrected p-distances*) slows down with time, eventually leveling off.
    A. This occurs because changes can reverse themselves, which is not accounted for by parsimony, leading to inconsistent estimation & long-branch attraction.
    B. Given a model of sequence evolution, we can obtain the probabilities of all of these possibilities and determine, after accounting for every possible path, what the most likely phylogeny is that could produce the observed data.
    C. Luca Cavalli-Sforza and Anthony Edwards in 1966, and later Felsenstein in 1973, introduced Maximum Likelihood methods for phylogeny reconstruction. However, it wasn't computationally tractable until Felsenstein's 1981 paper that introduced *the pruning algorithm*.

II. To consider a tree probabilistically, we need to remember three rules:
    A. The AND rule. The probability of two independent events is simply the product of the individual probabilities. For example, if you roll two die, then the probability that they both end up 6 is 1/6*1/6 = 1/36. For phylogenetics, we treat what happens on each branch as independent, since they represent separate lineages subject to independent evolutionary forces. Thus, to calculate the probability of the data being generated across the entire tree, we need to multiply the probabilities of separate events on each branch of the tree.
    B. The OR rule. The probability that two *mutually exclusive* events is simply the sum of the individual probabilities. For example, if you roll one die, the probability that it ends up 6 OR 1 is 1/6 + 1/6 = 1/3. For phylogenetic trees, this means that there are many sequences of events that can end up with a particular pattern of data. We must add up the probability of each possible pathway to generating a particular dataset to get the total likelihood.
    C. Conditional probability. A more general rule than the two above is the law of conditional probability. Namely that: P(A & B) = P(A) P(B | A) = P(B) P(A | B).
        1. If A & B are independent, this reduces to the AND rule: P(A & B) = P(A)P(B)
    D. With these 3 rules, we can calculate the probability of a particular generating process on a phylogenetic tree.

III. To obtain our likelihood, we will make a number of simplifying assumptions.
    A. Every column in our character matrix, or *site* in a *DNA alignment*, is independent. That is, we can calculate the likelihood of each individual site and simply multiply each of these likelihoods together to get the total likelihood of a full sequence.
        1. How do you feel about this assumption for sequence data? For morphology?
    B. Every branch in a phylogenetic tree is independent of every other branch.

C. The evolutionary process is *memoryless*. That is, as soon as a substitution occurs, all memory of past events is loss. Only the current state matters. This makes our model a *Markov model*.

1. How do you feel about this assumption for sequence data? For morphology?

D. At every instant of time, there is a *transition probability,* from one state to another. At any particular instant of time, only one thing can happen.

E. That the evolutionary process is *reversible* or *symmetric* with respect to time. These will apply to our molecular sequence evolution models, but not necessarily all of our morphological models. If the process is reversible, we can consider unrooted trees, as the direction of time does not matter, the probabilities are the same.

F. Together, we can put together these assumptions into a *Continuous-Time Markov Model*.

IV. Even the simplest phylogenetic models are complex models that have a number of parameters we must estimate. They are:

A. The topology of the tree, the branch lengths of every edge in the tree, the instantaneous transition probabilities, and the state frequencies.

B. We also have the unknown states of all unobserved nodes in the tree. These could be treated as parameters themselves. But what we would like to do is integrate over all possible states of these nodes. Thus, simply as a by-product of our modeling fitting, we will obtain for free the probabilities of *ancestral states* at each node of the phylogeny.

C. For example, for DNA data, we could imagine assigning every node one of the 4 bases (A, C, G, T) and calculating the probabilities, given our node states, parameters, tree and model, of ending up with our observed sequences. As a brute force approach, we could try every single combination of node states that are possible. So for a tree with 10 nodes, there would be 4^10 = 1,048,576 possible node states. This is a lot of probabilities to compute! Felsenstein’s pruning algorithm addresses this problem by realizing that many of the calculations across all those possibilities are highly redundant. Thus, we can break the calculation down into components by traversing down the tree and calculating small chunks at a time, keeping track of them, and combining them as we go. By doing so, we can get the full likelihood with a tiny fraction of the “brute-force” computation.

D. The pruning algorithm starts with a pair of sister species, calculates the likelihood of observing their two states given all possible node states, and saves the result. We then treat this node as a “tip”, and find the next pair of species down the tree. This “pruning” or “peeling” algorithm is so called because we find sister species pairs, calculate the likelihood, save the results for each state, replace the pair with a single tip, and repeat the process.

E. When we get to the root, we end up not being able to go any further. We have one last probability, the probability of A, C, G, and T at the root (for molecular data). These probabilities are not based on conditional probabilities, since they are the starting point. Instead, these will come from our *state frequencies* ($\boldsymbol{\pi}_A$, $\boldsymbol{\pi}_G$, $\boldsymbol{\pi}_T$, $\boldsymbol{\pi}_C$). Oftentimes, we use the equilibrium frequencies of the states given the model. Or you could put a “flat prior” on the root, so that all four bases were equal in probability.

V. Unlike parsimony, we are going to assume that the number of changes on a phylogeny are *time-dependent*. If there is more time, we expect more changes. If there is less time, we expect fewer or no changes. The expected number of changes is the product of the *substitution rate* and *time.* Since we have to estimate both of these parameters, they actually cannot be disentangled. They are *not identifiable*, in statistical jargon. Thus, we are going to estimate branch lengths in terms of the expected number of substitutions. Converting these branch lengths to units of time will require additional information, such as fossils.

A. We are going to model evolutionary change as a *Poisson process*. A Poisson process has a number of characteristics, including:

1. Changes are uniformly distributed across the tree
2. Changes have an exponential amount of time between them
3. The amount of time you have already been waiting has no influence on when the next event will occur (*memoryless* property).

B. Events occur at an instantaneous rate $\mu$. If you have a base A either it experiences and “event” (Paul Lewis calls these events “ACHNyons”, for “Anything Can Happen Now”) or it doesn’t. The probability that it doesn’t is:

$$P(\text{no event}) = e^{-\mu t}$$
$$P(\text{an event}) = 1 - e^{-\mu t}$$

C. So if we want to know the probability that A changes to A, we have two ways that can happen:

$P_{AA} = \ e^{-\mu t} + (1- e^{-\mu t}) * (1/4) = 1/4 + 3/4 * (e^{-\mu t})$

(the 1/4 comes from the probability that at least 1 event happens, but it ends up being substituted with an A again anyway).

D. Similarly, the probability of change is:

$P_{AT} = P_{AG} = P_{AC} = \ (1- e^{-\mu t}) * (1/4) = 1/4 - 1/4 * (e^{-\mu t})$

E. These “ACHNyons” are not necessarily substitutions. And we discussed that substitution rate is not identifiable from time. So we define the parameter $\nu$ as the expected number of substitutions.

$\nu = (3/4)\,\mu t$

$4\nu/3 = \mu t$

F. Thus we can formulate the model in terms of branch length ($\nu_i$):

$P_{AA} = \ 1/4 + 3/4 * (e^{-4\nu/3})$

$P_{AT} = P_{AG} = P_{AC} = 1/4 - 1/4 * (e^{-4\nu/3})$

We can use these probabilities to calculate the transition probabilities across branches on the tree. Then applying the pruning algorithm and combining with the root probabilities, we get the full likelihood.

VI. The above model is the simplest of the statistical phylogenetics models, the Jukes Cantor model. It is a Continuous-Time Markov Model, and is used in many applications besides phylogenetics. We can summarize the model with a tree topology, a vector of branch lengths, the state frequencies, and *the Q matrix.* Which is just an easy way to display the model in matrix form:

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -3\beta & \beta & \beta & \beta \\ C & \beta & -3\beta & \beta & \beta \\ G & \beta & \beta & -3\beta & \beta \\ T & \beta & \beta & \beta & -3\beta \end{array}$$

The diagonals are simply made so the rows and columns sum to 0. Each cell tells us the instantaneous rate we transition from (row) a base to (column) another base.

A. We can calculate transition probability matrix for a particular branch length $v$, by exponentiating the matrix. This is by far the most computationally intensive part of fitting this model.

$$P = e^{-Qt}$$

B. Modifications of the Q matrix will be the basis of a HUGE number of models we will learn in class. The simplest JC69 (Jukes Cantor 1969) model has only 1 parameter, and is time-reversible, as seen by the symmetry of the matrix. This imposes a lot of assumptions on our model of trait/nucleotide evolution. Many of these assumptions are VERY BAD. We thus have room for improvement.