Substitution models and discrete character models I

I. Last time we learned the most basic model used in statistical phylogenetics, the Jukes Cantor model (JC69, phylogenetics models are usually named with the initials of the authors followed by the year).

II. What are the assumptions of JC69? We can enumerate 7 very important assumptions. Think about how biology can break these assumptions. Can we modify the model to address these?
   A. All substitutions are equally likely
   B. Base frequencies are equal
   C. Every site has equal probability of substitution
   D. Process is constant through time and across branches
   E. Sites are independent of each other
   F. Substitution is Markovian (memoryless)
   G. All sites have the same evolutionary history

III. All substitutions are equally likely
   A. We know the above assumption to often be extremely wrong. Luckily, we can easily modify the Q matrix to give unique parameters to Transversions vs. transitions (as is done in the Kimura 2-Parameter K2P/K80 model). At the most extreme case, we can give unique rates to each transition combination, leading to 6 rates. This is the General Time-Reversible model (GTR). If we are Ok searching rooted trees, we can make the model even more general by accommodating asymmetric rates.

IV. Base frequencies are equal
   A. Again, we can modify the vector of state frequencies to accommodate unequal base frequencies. We can either 1) set base frequencies to equal 2) estimate the equilibrium base frequencies under the best-fitting model or 3) set the base frequencies to their empirically observed values.

V. Every site has an equal probability of substitution
   A. Relaxing this assumption can often result in huge increases in likelihood. For example, in protein coding genes we know that the third codon position changes much more rapidly than other sites. Thus, we can partition our dataset into two parts. In one, we put 1st and 2nd codon positions. In the other, we put 3rd codon positions. We essentially fit two different models to these two different partitions. This is the “site-specific approach”.
      1. Site-specific approaches usually use *a priori* designations of partitions. For example, you can partition by gene, and then within each gene, partition by codon position. Or for ribosomal RNA genes, you could partition by stem and loop, for example.

2. The drawback of site-specific approaches is that you have to know the partitions *a priori.* (Though there are approaches being developed to automate this process and integrate it into the phylogenetic estimation procedure). This means that often, we will put several sites in a particular partition that may not evolve by a common model. For example, one particular site may be under positive selection in a protein, while another is under purifying selection.

B. Alternatively, we can propose a mixture model. One common mixture model is adding "invariant sites", while another is "gamma-distributed rates". Both can be combined into a single model as well.
    1. Invariant sites suppose that there are a proportion of sites that never evolve at all, P_{invar}. This proportion is given a weight, and for every site, you calculate two likelihoods. One is the probability of observing the data for that site if it were invariant. The other is the probability of observing the data under the standard CTMP being estimated. If the site is variable, then it clearly isn't invariant and that likelihood will be 0 (essentially, it will be thrown away). Alternatively, if the site is fixed, it could either be that it is an invariant site, or it could be a variable site that just so happens to be constant by chance. This then allows the model to account for highly unlikely invariant sites under the CTMP by considering those likely to be invariant sites.
    2. We can also define a continuous function that assigns rates to low and high categories. A commonly used function is the discrete gamma distribution. We can set a certain number of categories (e.g. 4) that bin rates together based on this gamma function. We take the average rate for each bin by dividing the gamma function into 4 bins with equal area under the curve (density) and set the rates to be the average of that bin. The likelihood for every site is calculated under each of those 4 possible bins. We can estimate the alpha parameter during the analysis, and thereby estimate the degree of rate heterogeneity in our dataset.
    3. The main difference between the mixture and site-specific approach is that the mixture model lets each site "decide" which rate category it wants to be in, since each site is fit under each of the rate categories. The likelihoods are combined, so technically it is fit to all of the categories at once, but whichever one "fits better" is the one that will dominate the resulting likelihood. But this also means it's more computationally expensive and there are more parameters.
    4. Combining invariant and gamma distributed rates ("+I+G") is possible, but also can be dangerous. These will often result in flat likelihood surfaces, low identifiability of parameters, and poorly behaved statistical properties if there is insufficient data in the dataset.

VI. Process is constant through time and across branches

A. By and large, this is accommodated by the estimation of branch lengths. The tree is estimated with branch lengths in terms of expected substitutions per site, rather than time. Thus, combining the resulting tree with fossil calibrations or other constraints will reveal where on the tree rates have increased or decreased. However, it is also possible to have a model in which the distribution of rates follows a specific distribution or process. These models become much more important when we get to the topic of divergence time estimation using fossils and other constraints.

VII. *Sites are independent of each other*

A. Within our world of CTMP, we can't do much about this one. We can group sites together, but doing so greatly increases the number of parameters. For example:

1. Imagine two sites that can each take two states. We could define a 4 state character for each of the two states. This gives 00, 01, 10, and 11 as the possible states. We then could evolve between those states. This allows correlation because if the rate of transition from 00 -> 01 is not the same as the rate of transition of 10 -> 11, then that means they are evolving in a correlated manner. Unfortunately, even assuming only 1 trait changes at a time, this results in 8 transition rates for two characters with 2 rates. The number of parameters in these models quickly explodes as we increase the number of states. This makes creating these large multistate characters impractical for most datasets beyond a handful of characters.

VIII. *Substitution is Markovian*

A. The best way to deal with this is to transition to another kind of model we will talk about, the Threshold model. However, this model has not been implemented for phylogeny estimation.

IX. *All sites have the same evolutionary history*

A. This assumption will lead us to species tree estimation, rather than gene tree estimation. Whenever you see a paper say they analyzed a *concatenated dataset*, they are assuming that every gene has an identical evolutionary history. One major way this assumption breaks down is with incomplete lineage sorting. Another is hybridization. We will explore this issue in more detail later on in the course.