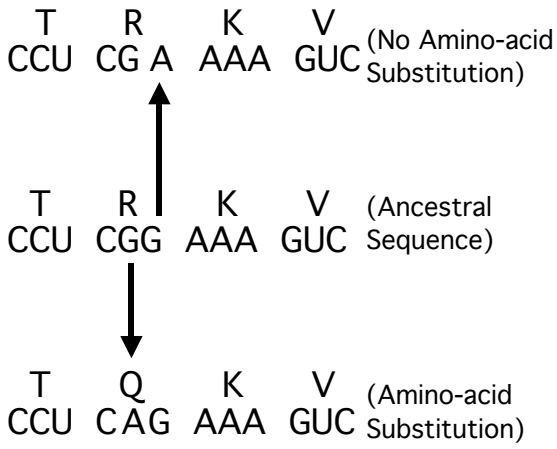
Lecture 11 – Increasing Model Complexity

I. Introduction. At this point, we've increased the complexity of models of substitution considerably, but we're still left with the assumption that rates are uniform across sites.

Differences in functional and structural constraints across sites leads to different sites evolving at different rates.

For example, if we look at this hypothetical sequence, we'll see that, because of the nature of the genetic code, not all nucleotide substitutions will result in an amino-acid substitution.



Therefore, a very common observation is that sites at 3rd -codon positions evolve fastest (perhaps at the rate of neutral mutation), followed by those at 1st positions, and 2nd position sites evolve the slowest.

This is one form of among-site rate variation, which exacerbates the loss of historical information caused by multiple hits.

If we think about this, it should be pretty obvious. If we have 10 substitutions, and they are distributed randomly across 50 sites, there should only rarely be more than a single substitution per site. However, if those 10 substitutions are distributed across 50 sites in a non-random fashion, say concentrated to 1/3 of them, many more will occur at multiply hit sites.

Because of the importance of this, I want to present ways to model among-site rate variation.

II. Discrete Methods:

The simplest thing to do is to assign the sites of an alignment to a series of rate partitions. This assignment is often done based on some extraneous information such as codon structure or stem/loop structure.

Accommodating different rates of substitution is easily accomplished simply by adding a relative-rate parameter *r* for site classes to our models, as illustrated below for a JC model:

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

Then the likelihood for a particular site is calculated as follows:

$$\ln L_{\tau}^{i} = \sum_{r=1}^{c} w_r \ln L_{(i,r)}$$

where, there are *c* rate categories, and w_r is the probability that site *i* belongs to a particular rate category; these are binary (0 or 1) if we're assigning sites to rate classes *a priori*.

B. The earliest common discrete-rates model is called the **Site-Specific Rates (or SSR) model**.

In the SSR models, the theoretical limit to the number of rate categories is the number of sites in the alignment, but usually these are determined *a priori* and often they follow codon structure.

Felsenstein gives an example of these on **page 223** in the text.

So, in this case, w_1, w_2, and w_3 are fixed to 0 or 1, and we just use an independent JC model for each class (e.g., codon position). The relative rate parameters then can be assigned, as Joe describes in the text, or they can be optimized numerically, which is what is often done.

These SSR models have the advantage that one can essentially use a different transformation matrix (**Q**) for each class. This can lead to huge improvements in fit between model and data relative to, say, a single GTR for all sites.

They have the disadvantage that all sites within a category are assumed to be evolving a uniform rate. This may be an ok assumption for 3rd codon position sites, but it's probably a really bad one for 1st and 2nd position sites (e.g., Buckley et al. 2001. Syst. Biol. 50:67).

C. Invariable Sites model.

Another common approach allows two rate categories and in one of these, the relative rate parameter is zero. This is based on observations that there are sites in alignments of conserved genes in which all of life has the same state.

We can think about this model in two ways.

As a mixture model.

Typically, in this model, the w_r for the category of sites that is potentially variable is estimated from the data. This is taken as the probability that a particular site belongs to either the variable class or the invariable class. So, the degree of the mixture is 2.

$w_{invar} = p_{invar}$	The probability that a site in in the class where $r = 0$.
$w_{var} = p_{var}$	The probability that the site is in the class where $r \neq 0$.
$w_{var} = 1 - w_{invar}$	Sites that are observed to vary have $w_{invar} = 0$.

We can also take the w_r to be the proportion of sites that are invariable across the alignment.

This model, then, is governed by the parameter p_{invar}, the proportion of sites that are invariable.

This is constrained to be $\leq$ the proportion of sites that are constant because there is a non-zero probability that a potentially variable site has not experienced a substitution due to stochasticity of the process (be careful with the term "invariant sites").

II. Continuous Methods:

There's no biological reason, however, to expect rates to fall into discrete categories, and we can use rate-mixture models to deal with this.

There are a number of continuous-rates models that have been applied historically, but one of these has become pretty dominant.

Almost all studies that attempt to incorporate ASRV directly use a gamma distribution to model rate variation across sites (**Γ-distributed Rates model**)

Gamma distributions are governed by two parameters: a shape parameter (α) and a scale parameter (β). The mean of a Γ-distribution is equal to the product of these, $\alpha\beta$.

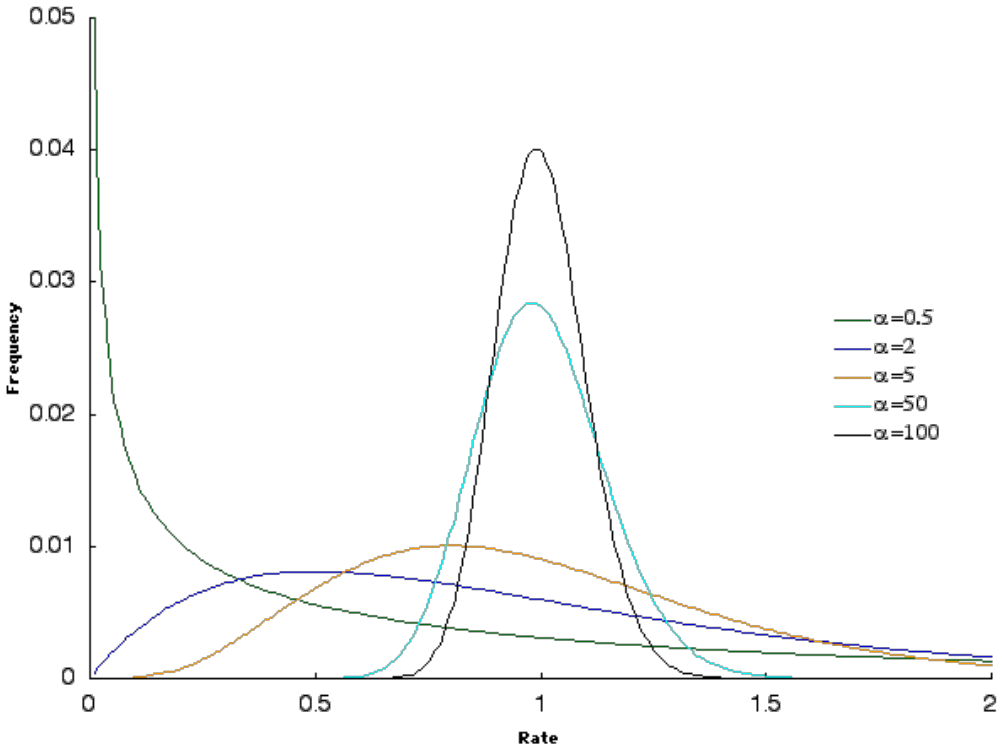
In applications to molecular systematics, we set the mean of the Γ-distribution equal to 1 by constraining $\beta = 1/\alpha$.

This allows us to scale branch lengths in units of expected substitutions per site. In addition, the Γ-distribution is then governed solely by the shape parameter.

The advantage of using Γ-distributions to model ASRV is that, by varying this single parameter, α, the distribution can take on a variety of different shapes.

When $\alpha = \infty$, the gamma model converges on a single rate model.

When $\alpha = 0.5$, the distribution becomes L-shaped.



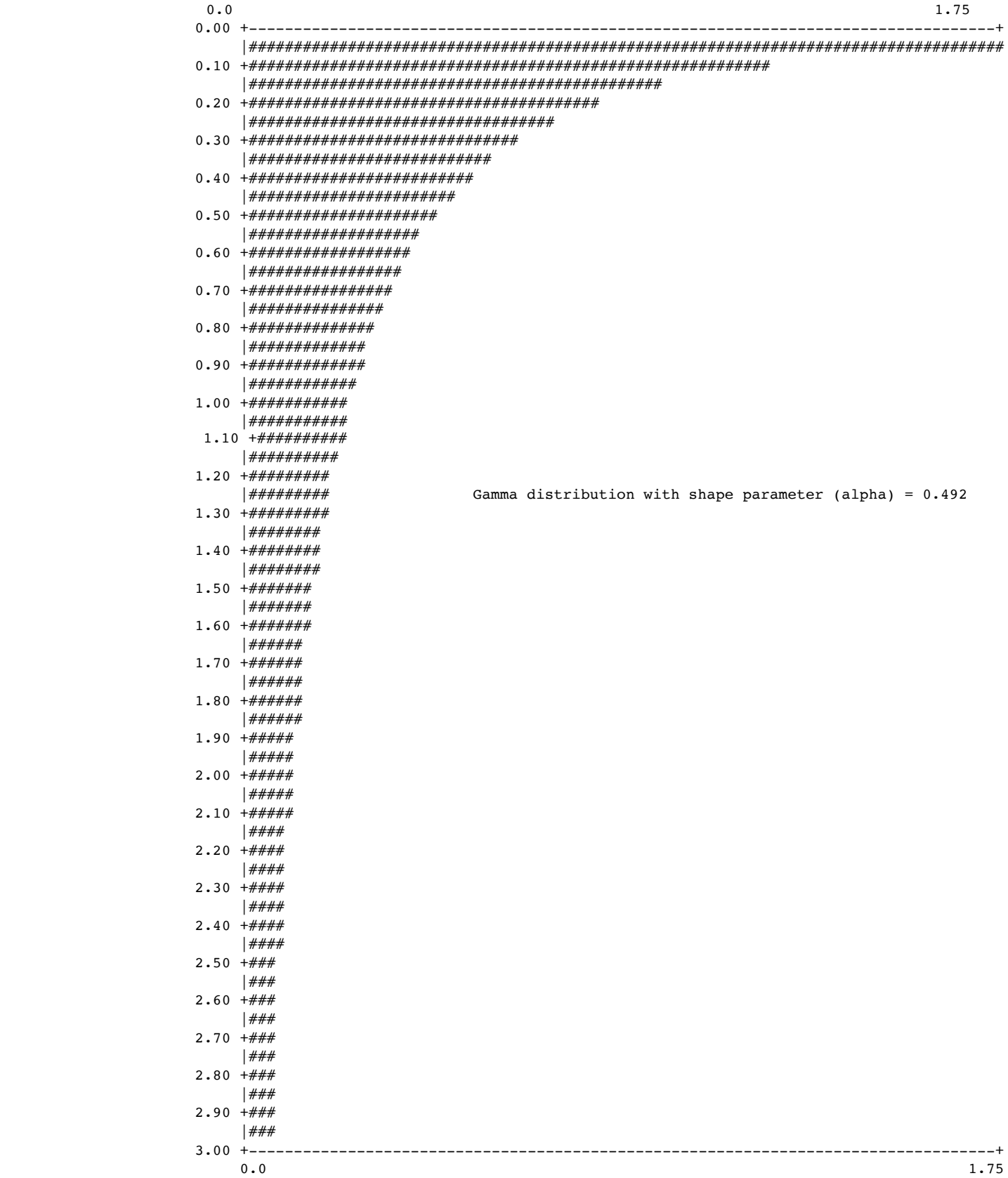
Many real data sets have a shape parameter of ≈ 0.5, although there is a lot of variation.

So, using a gamma distribution to model ASRV in phylogenetics can be accomplished by integrating across the Γ-distribution. This isn't at all feasible, so the common solution is to discretize the gamma distribution.

The idea is to break the continuous gamma into a number of rate categories (usually 4 – 8). The rate within a category is represented by the within-category mean, and these means are drawn from a Γ-distribution with shape parameter α (Yang, 1994. *J. Mol. Evol.* 39:306).

The boundaries of the rate categories are set such that there is an equal area of the distribution in each.

This is demonstrated below:


0.0
1.75
Gamma distribution with shape parameter (alpha) = 0.492

Cut-points and category rates for discrete gamma approximation (ncat = 4)

category	------ cut-points ------ lower	upper	rate (mean)
1	0.00000000	0.09804816	0.03191473
2	0.09804816	0.44841399	0.24666120
3	0.44841399	1.31969682	0.81435904
4	1.31969682	infinity	2.90706503
			Mean = 1.0

So, these means are incorporated into the likelihood function as the r_i's, and the w_i's for each site (the probability of the site occurring in rat category *i*) are optimized.

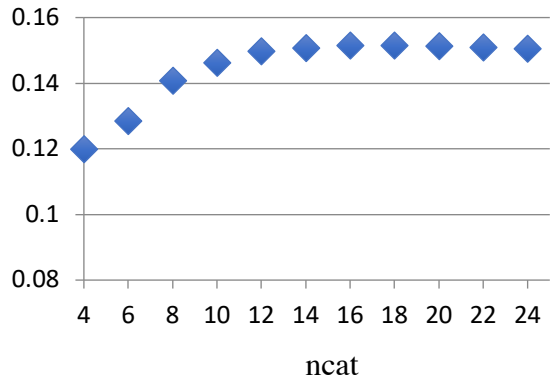
Each site has a non-binary set of w_i's, so this is actually a mixture model.

Although it's not usually done, one should vary ncat and identify the smallest value that produces an accurate estimate of α. *This is data-set dependent*.

Cut-points and category rates for discrete gamma approximation (ncat = 8)

category	------ cut-points ------ lower	upper	rate (mean)
1	0.00000000	0.02338747	0.00768838
2	0.02338747	0.09804816	0.05614108
3	0.09804816	0.23352213	0.16013076
4	0.23352213	0.44841399	0.33319164
5	0.44841399	0.78071211	0.60229167
6	0.78071211	1.31969682	1.02642641
7	1.31969682	2.35886822	1.77009489
8	2.35886822	infinity	4.04403517

and:



This is done across the entire data set, so essentially, we take the same transformation matrix (**Q**) for each site and **scale it by the average rate for each category**.

This has the tremendous advantage of being able to accommodate such a wide diversity of rates with just a single parameter, α. Some sites can be so slowly evolving to have a high probability of stasis, yet others (perhaps adjacent) may be free to evolve rapidly.

It has the disadvantage that we apply the same transformation matrix (**Q** matrix) uniformly across a data set.

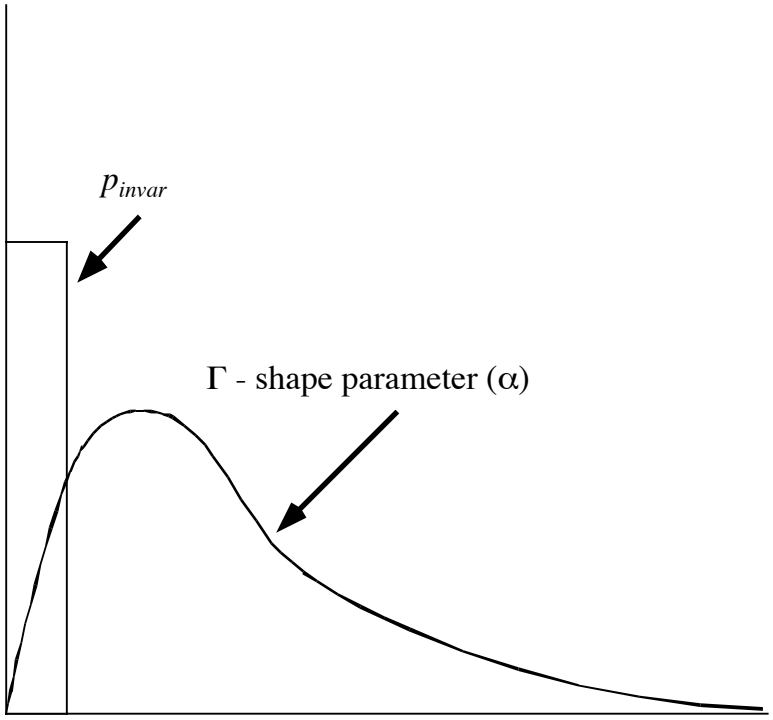
We may get a better fit if we allow **Q** to be different for each category.

It's also pretty common to over discretize the gamma distribution (i.e., use too few rate categories).

SSL's are calculated ncat times, and this represents a very constrained mixture model.

III. I+G Models

A further elaboration that has become widely used is a mixture of invariable sites, with rates at variable sites being drawn from a gamma distribution. This is called the **I + Γ model**, and was developed independently by Gu, Fu, & Li (1995, Mol. Biol. Evol.) and Waddell and Penny (1996).

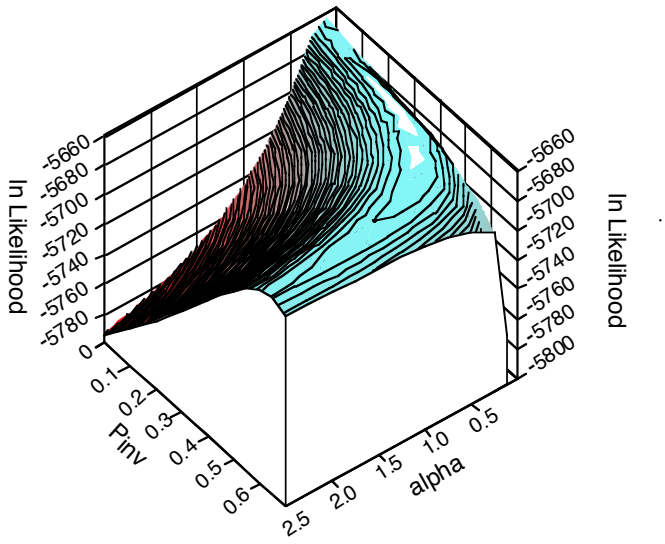


This is intuitively very appealing when one considers that, at least from some genes, there's a set of sites that are constant across essentially the tree of life.

This model is very frequently required by real data sets (as assessed by methods we'll discuss in subsequent lectures), but there are some issues with it that are sometimes not appreciated.

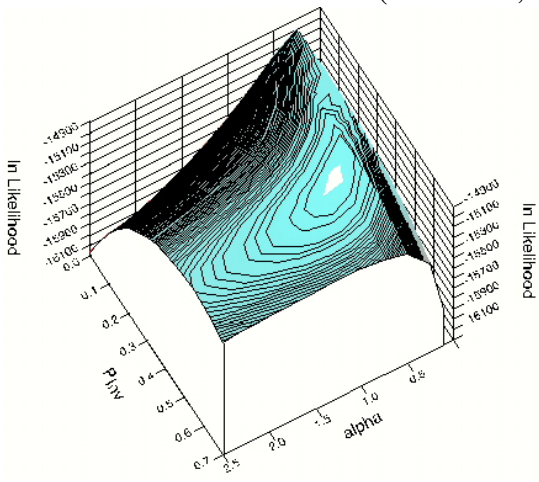
The mixed model and the gamma alone expect there to be many constant sites. It can be very difficult to discern the sites that are truly invariable from those potentially variable sites that are evolving slowly enough to have a high probability of stasis.

This can result in very poorly behaved likelihood surfaces (non-identifiability), as shown below (from Sullivan et al. 1999. Mol. Biol. Evol., 16:1347-1356).



This is the likelihood surface for the parameters of the I + Γ model. There are relatively few taxa in this data set and there are multiple peaks in the likelihood surface, one of which is the true peak (the data are simulated, so fit the model perfectly).

However, with many taxa, the surface is better behaved (same data, more taxa).

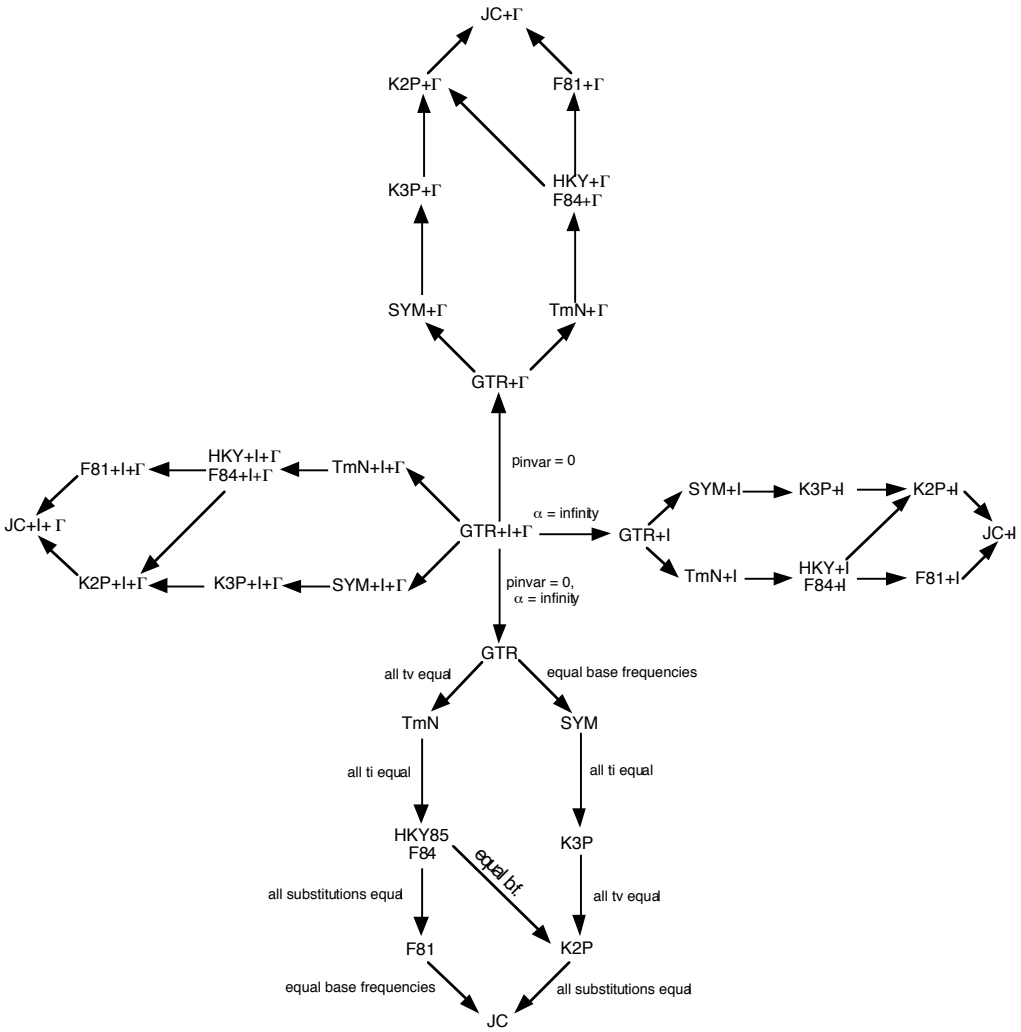


IV. Expanding the GTR family of models.

Let's think about the parameters of a GTR transformation matrix with ASRV modeled with I + Γ. There are 10 free parameters. Eight are associated with the transformation matrix: three free base frequencies (they're constrained to sum to 1) & five relative rate parameters (they're relative and r_{GT} is set to one). Two are associated with ASRV: the shape parameter of the gamma distribution and p_{invar}.

Just as, when we were discussing the transformations matrix, simpler models are special cases of the most parameter rich, the equal rates models are special cases of the variable rates models. That is, we can erect a nested series of substitution models.

So, the relationships among the GTR+I+Γ family of models can be illustrated with a clover-leaf diagram. To me, this is a very convenient way to visualize model space.



In the I+Γ models, if α = infinity, the ASRV model is an invariable sites model. If $p_{invar} = 0$, the ASRV model is equivalent to a Γ alone.

In a Γ model alone, when α = infinity, the gamma model converges to the equal-rates models. Similarly, in an invariable-sites model (alone) if $p_{invar} = 0$, the invariable sites model also reduces to an equal-rates models.

Remember that each lobe of the cloverleaf represents 203 possible restrictions of the r-matrix.

Similarly, we can consider the SSR models to be a family of special cases. If we have a GTR+SSR_3 model, we can think of the following parameterization:

π_{A1}	π_{A2}	π_{A3}
π_{C1}	π_{C2}	π_{C3}
π_{G1}	π_{G2}	π_{G3}
π_{T1}	π_{T2}	π_{T3}
$r_{(AC)1}$	$r_{(AC)2}$	$r_{(AC)3}$
$r_{(AG)1}$	$r_{(AG)2}$	$r_{(AG)3}$
$r_{(AT)1}$	$r_{(AT)2}$	$r_{(AT)3}$
$r_{(CG)1}$	$r_{(CG)2}$	$r_{(CG)3}$
$r_{(CT)1}$	$r_{(CT)2}$	$r_{(CT)3}$
$r_{(GT)1}$	$r_{(GT)2}$	$r_{(GT)3}$

The GTR model applied to all sites is equivalent to this with the following restrictions:

π_{A1}	=	π_{A2}	=	π_{A3}
π_{C1}	=	π_{C2}	=	π_{C3}
π_{G1}	=	π_{G2}	=	π_{G3}
π_{T1}	=	π_{T2}	=	π_{T3}
$r_{(AC)1}$	=	$r_{(AC)2}$	=	$r_{(AC)3}$
$r_{(AG)1}$	=	$r_{(AG)2}$	=	$r_{(AG)3}$
$r_{(AT)1}$	=	$r_{(AT)2}$	=	$r_{(AT)3}$
$r_{(CG)1}$	=	$r_{(CG)2}$	=	$r_{(CG)3}$
$r_{(CT)1}$	=	$r_{(CT)2}$	=	$r_{(CT)3}$
$r_{(GT)1}$	=	$r_{(GT)2}$	=	$r_{(GT)3}$

So, we have a number of models, and there are nested series.

GTR+CAT in RAxML (and FreeRates Model in IQ-TREE)

Before we leave rate-heterogeneity, we should discuss different approach relaxes the assumption that rates conform to a Gamma distribution.

It's like the SSR approach, in that sites are assigned to rate classes, and therefore the w_r's are all either zero or 1.

However, sites are classed into categories (usually 25 rate categories) based on an initial estimate their rates on a starting tree.

Individual relative rates are estimated for each site.

The rates for each class (r) are assigned as the rate of the site with the highest SSL in the category, and they're then fixed for tree searching (remember via stepwise addition under parsimony followed by lazy SPR).

This is faster than a Γ-model because SSLs are only calculated once (instead of *ncat* times), since every site is *assigned* to a single rate category.

This approach works well when the number of sequences in the data set is large, but when there's less than several hundred, the estimates of the rates at each site are pretty lousy and the performance of GTR+CAT declines.

IV. rRNA Model

A couple models have been developed to deal with non-independence of nucleotides in paired stem region of rRNA.

These models use *a priori* partitioning of sites into stem and loop regions, and sites in the loops partition are treated with some variant of the GTR+I+Γ family.

Sites in the stem regions are treated using the doublet model.

Doublets are treated as characters rather than nucleotides and there are 16 possible states rather than 4.

So instead of 12 substitution types (or 6 reversible types) there are n(n-1) = 240 types (or 120 reversible types).

In addition, instead of three free base frequency parameters, there will be 15 free doublet frequencies, many of which are likely to be zero.

Smith et al. (2004. Mol. Bio. Evol. 21:419) used an aligned database of 50K sequences to estimate these parameters and they provide a fully parameterized empirical model.

V. Codon-based models

It's also possible to model the non-independence of sites generated by the genetic code using codon-based model.

Here, in-frame triplets are used as characters and there are 61 possible character states (64 triplets minus the three stop codons). Thus, the transformation matrix has 3660 rate parameters (or 1830 in the reversible case).

Again, empirical matrices can be used.

Alternatively, cells of the transformation matrix can be restricted so that there are only, say two substitution types.

TTT ←→ TTC: Both code for Phe so the T←→C transition is silent. The cell in the matrix would be filled with $\alpha\pi_C$, where α is the rate of silent substitutions and π_C is (as before) the frequency of nucleotide C.

Conversely, TTT←→ TTA would be expressed as $\beta\pi_A$, where β is the rate of amino acid replacement substitutions, because TTA codes for Leucine.

This is the approach taken by Muse & Weir (1994. Mol. Biol. Evol. 11:715). There are only 4 parameters here, the three free base frequencies and the ratio of the rates of silent vs. replacement substitutions.

Goldman and Yang (1994. Mol. Bio. Evol., 11:725) go a step further and incorporate a transition/transversion rate ratio, and Halpern and Bruno (1998. Mol. Biol. Evol. 15:910) allow all six possible nucleotide substitution types.

A cool thing about this approach is that we can calculate the ratio of synonymous to replacement substitutions, which allows for an assessment of the strength of selection operating at a site.

Take a minute and make sure we have our usage of “relative rates” straight.