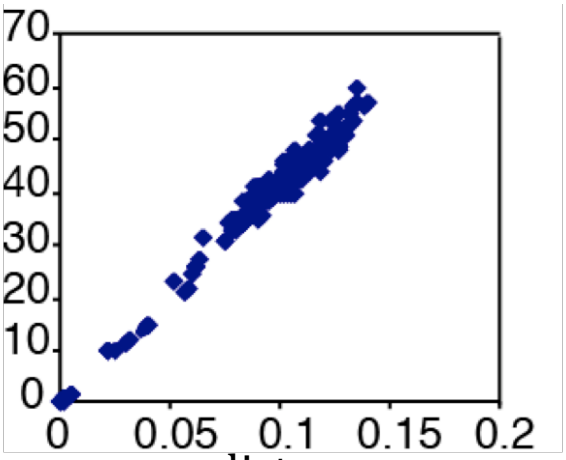
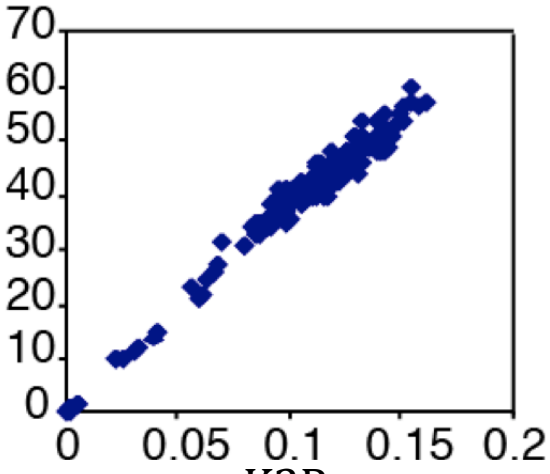
Lecture 12 – Selecting Models

I. Introduction. So, we have an array of models to choose from. Now the goal of phylogeny estimation isn't simply restricted to choosing a tree topology, but also to find the optimal combination of $2n$-3 branch lengths, model, and model parameters, in addition. Felsenstein presents this material in the first half of Chapter 19. There's a good (but slightly dated) review in Sullivan and Joyce (2005. Ann. Rev. Ecol. Evol. Syst., 36:445) and another from a slightly different perspective in Posada and Buckley (2004. Syst. Biol., 53:793).

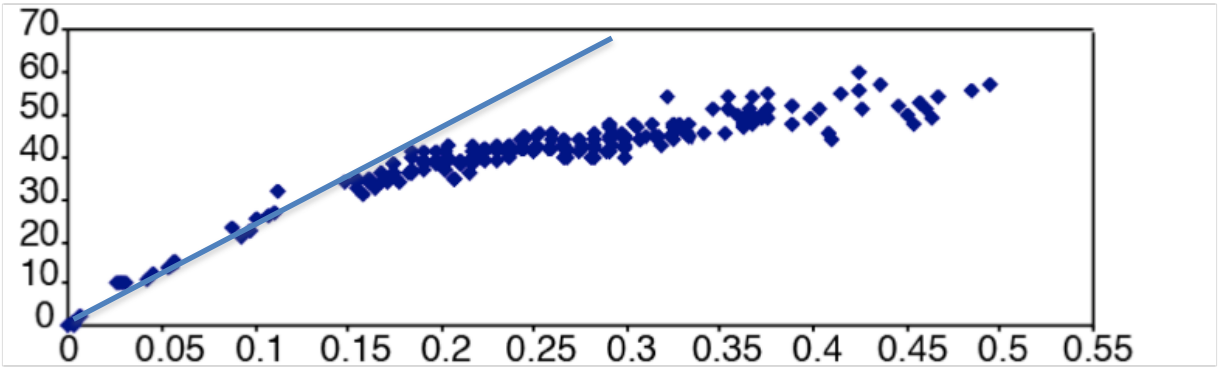
All models are wrong, but some are useful.



p-distance



K2P



HKY+I+Γ

All these are from the exact same data (Cicero & Johnson 2001). The authors selected HKY+I+Γ for ML analyses. Lots of folks used to use K2P for saturation plots; it's both wrong and misleading for detecting multiple hits. The HKY+I+Γ model is wrong (surely) but is nevertheless useful. Even for data exploration, model choice can matter.

There are several approaches to selecting a model for phylogeny estimation. There are those based on absolute goodness of fit, relative goodness of fit, and more innovative approaches that incorporate both fit and performance.

All these are based on the fact that the likelihood score is interpretable as a measure of fit between the model and data and that this is directly comparable across models (contrast this with tree length across weighting schemes under parsimony).

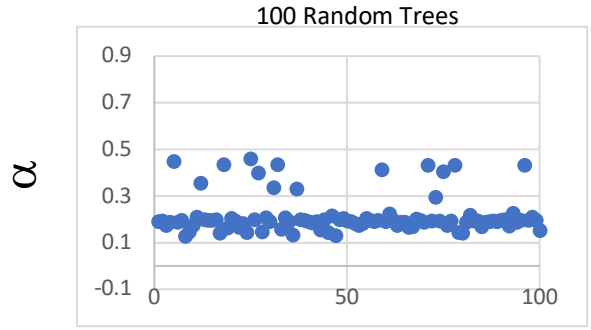
The first thing to note is that most approaches to model-based phylogeny estimation require that the model be selected and defined prior to any tree search.

A. Iterative Approach.

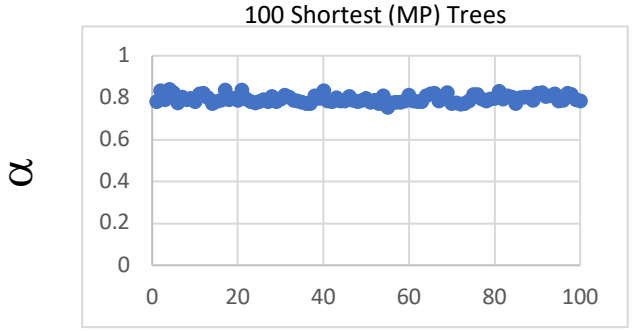
Ideally, we could evaluate models (and estimate the parameters of those models) and topologies simultaneously.

One way around that that has been widely used is based on understanding of the manner in which parameters vary across topologies. (This was the topic of chapter 2 of my dissertation: Sullivan, J., K. E. Holsinger, and C. Simon. 1996. The effect of topology on estimates of among-site rate variation. *Journal of Molecular Evolution*, 42:308-312.)

Parameters estimated from random trees can exhibit a strong bias and can be very poor estimates. Here, data are simulated with $\alpha = 0.8$.



But even decent trees can provide better estimates than random trees:



So, any decent tree (i.e., better than random) actually provides decent estimates of model parameters; they are not terribly different than those derived from the ML tree.

This is from the data set you're using in lab. As you can see, even though the NJ tree is substantially worse than the ML tree (ca. 5.3 ln L units), the MLE's of parameters (of the HKY+I+Γ model) are not very different when optimized on each.

Tree	ML	NJ
-ln L	6591.293	6586.008
Ti/tv:		
exp. ratio	2.684860	2.854459
kappa	5.503353	5.850992
Shape	0.787432	0.766300
P_inv	0.456164	0.456575

This suggests that we ought to be able to use a rapidly built approximate tree (i.e., an NJ tree or an MP tree) on which to estimate model parameters and select a model.

So, this is a successive approach commonly implemented.

Step 1: Construct an initial tree using NJ with LogDet distances. Save the tree.

Step 2: Calculate the likelihood score of alternative models of nucleotide substitution, with all model parameters optimized simultaneously. Choose from among these model based on some (statistical) criterion.

Step 3: Conduct a new search of tree space under the likelihood criterion, using the model chosen in Step 2, with the parameters of that model fixed to values estimated in Step 2. If the trees found in the current search are a subset of the trees found in the preceding search, STOP; otherwise, go back to Step 2.

Again, this has become the standard manner in which ML trees are estimated, and it seems to work pretty well (Sullivan et al., 2005. Mol. Biol. Evol. 22:1386).

Let's take a look at the stage here where we evaluate alternative models. There are a few different approaches that have been applied to model selection in phylogenetics, and we'll discuss some of them.

II. Absolute Goodness of Fit.

Perhaps the most intuitively appealing approach to model selection would be to assess the absolute goodness of fit between model and data.

This is based on the expectation that the model that fits the best should perform the best.

So how does one assess goodness of fit in an absolute sense? It is actually doable, but it is rarely done.

As we discussed Monday, the test was first developed by Nick Goldman in 1993 (J. Mol. Evol. 36:182 – it's called the Goldman-Cox test) and is an extension of the parametric bootstrap.

In order to understand this test, we need to introduce the concept of the unconstrained likelihood or the maximum possible likelihood that a data set could possibly achieve.

Remember that we only keep track of site patterns, and the frequency with which they occur.

```
   1       A G T A C A . . . . . . . . . . . . . . . .
   2       A G T A . . . . . . . . . . . . . . . . . .
   3       A G T A . . . . . . . . . . . . . . . . . .
   .       . . . . . . . . . . . . . . . . . . . . . .
   .       . . . . . . . . . . . . . . . . . . . . . .
   n       A G T A . . . . . . . . . . . . . . . . . .
Pattern    1 2 3 1 . . . . . . . . . . . . . . . . .(a)
```

such that:

$$\ln L(\tau) = \sum_{a=1}^{4^n} f_a (\ln L_a(\tau))$$

Now this value has an upper bound. It's bounded by the condition in which the model predicts the data exactly.

That is, when:

$$L_a = f_a.$$

Thus, the upper bound is given by:

$$\ln L_{\max} = \sum_{a=1}^{4^n} f_a (\ln f_a)$$

So, this is simply calculated from the histogram of the site patterns in the data. This can be thought of as the likelihood-score we would achieve if every site had its own model and could evolve on its own tree.

We can measure the deterioration in likelihood score associated with forcing all the data to fit a single model and tree. This quantity is given by:

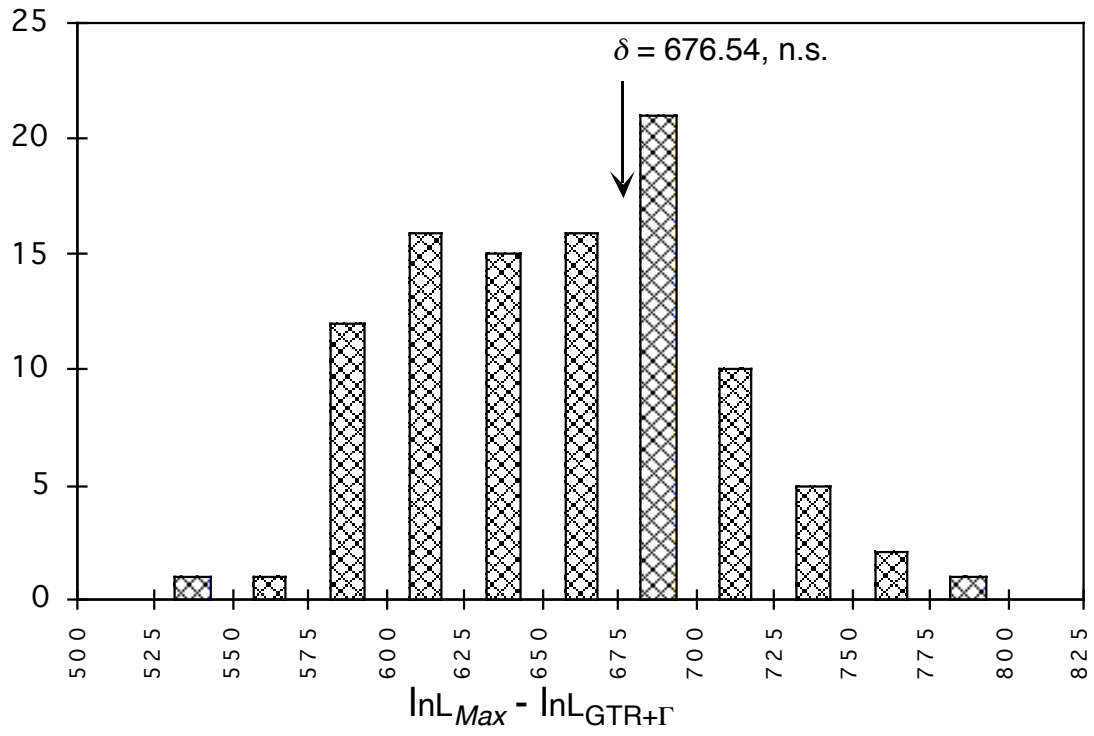
$$\delta = \ln L_{max} - \ln L_{(\tau|\text{Data,Model})}$$

So, now we have a test statistic that measures how well the data fit the model, and we need to find a way to assess its significance.

This is where the parametric bootstrap comes in. We have this quantity as we've estimated from the real data, where we don't know how good the model is. Now we can simulate the distribution under the null hypothesis of a perfect fit between model and data.

In the simulated data sets, we know that there's a perfect fit between model and data because we're using the model under examination to generate the test statistic in the replicate.

Such a test is shown here for a data set I published several years ago.



The test statistic for the real data was 676.54 ln*L* units. That falls right in the middle of the distribution I generated under the null hypothesis of a perfect fit.

So, in theory, one might be able to do such a test on all models under consideration and pick the simplest one that is not rejectable.

The logic for this is that a more complex model will necessarily have a better likelihood score than a simpler model, even if the simpler model is true. This is because **the extra parameters in the more general model will explain stochastic variation, even if they're superfluous.**

Remember though that even if we simply restrict our attention to the GTR+I+Γ family of models, there are a prohibitively large number of models we might be interested in considering and we would have to construct the null distribution for each via simulation to select the simplest model that exhibits an adequate fit.

An alternative approach is through use of posterior predictive simulations (Bollback, 2002. Mol. Biol. Evol., 19:1171), which incorporates uncertainty in parameter estimation better.

We therefore need a method of evaluating models prior to conducting intensive data analysis, especially because these seem to be very low power tests (Ripplinger and Sullivan, 2010. Mol Biol. Evol., 27:2790). Jeremy Brown has developed such tests (e.g., Brown. 2014, Syst Biol. 63:334) that are inference based.

III. Relative Goodness of Fit

Given that it's not feasible to do a series of *absolute* goodness-of-fit tests, most methods of *a priori* model selection have focused on the *relative* goodness-of-fit of alternative models.

That is, we can evaluate the relative fit of each of the set of models that we're interested in testing and choose one on this basis.

The first issue to deal with is what measure of fit we should be using? There are a number of alternatives.

A. Akaike Information Criterion (AIC).

The AIC was developed in 1974 to penalize models that are over-parameterized.

$$\mathrm{AIC}_i = -2 \ln L_i + 2d_i,$$

where d_i is the number of parameters in model i. This can be calculated for all models being considered and the one with the minimum AIC can be chosen.

Thus, the AIC includes a measure of fit (the likelihood score of the model) and a penalty for over-parameterization.

The AIC_C corrects for small sample size by multiplying the by, $2d_i$, by: $(2d_i^2 + 2d_i) / (n - d_i - 1)$, where n is the sample size (approximated by number of sites).

This converges to the traditional AIC with data sets of the size typically used in phylogenetics; $n \sim 400$ (unless you're examining highly partitioned models – more on these later).

The theoretical justification for this is that it is an approximation for the Kullback-Leibler distance:

$$E\,(\ln P(D \mid M_T) \mid M_T) - E\,(\ln P(D \mid M_i) \mid M_T),$$

Where M_T is the true model and M_i is the model under consideration. So, the AIC minimizes the information lost by using the approximating model *i* relative to the information in the (unknown) true model, *T*.

B. Bayesian Information Criterion (BIC).

A similar measure also includes information on the sequence length (at least when it's applied to molecular phylogenetics).

$$\mathrm{BIC}_i = -2 \ln L_i + d_i \ln(n),$$

where n = the number of observations (usually the number of sites, but this is approximate).

The BIC, like the AIC, includes a measure of fit (the likelihood score of the model) and a penalty for over-parameterization, and the BIC penalizes over-parameterization more than does the AIC.

This actually approximates the probability of the model given the data and tree.

Again, the BIC can be calculated for all models being considered and the one with the minimum BIC is equivalent to model that has the highest probability under certain assumptions.

1. A uniform, or at least "sufficiently vague" distribution of priors across models.

2. The Taylor approximation holds; the joint likelihood is a good approximation of the marginal likelihood.

Evans and Sullivan (2011. MBE. 28:343) demonstrated that this is the case as long as there is a lot of information in the data regarding model preference (i.e., there are few models in the 95% credibility interval of the posterior).

C. Hierarchical Likelihood Ratio Tests (LRTs)

LRT's are a general statistical method designed for testing model assumptions. In general, they are restricted to comparisons of a pair of nested models.

They were the first model-comparison tests used in molecular phylogenetics.

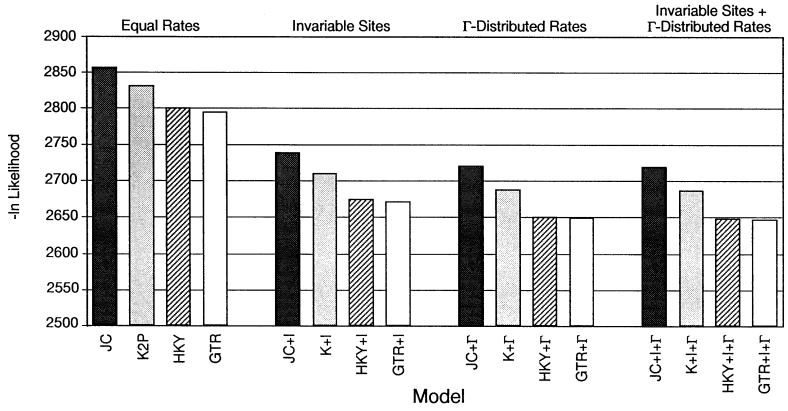
The test statistic is:

$$\delta = 2(\ln L_0 - \ln L_1)$$

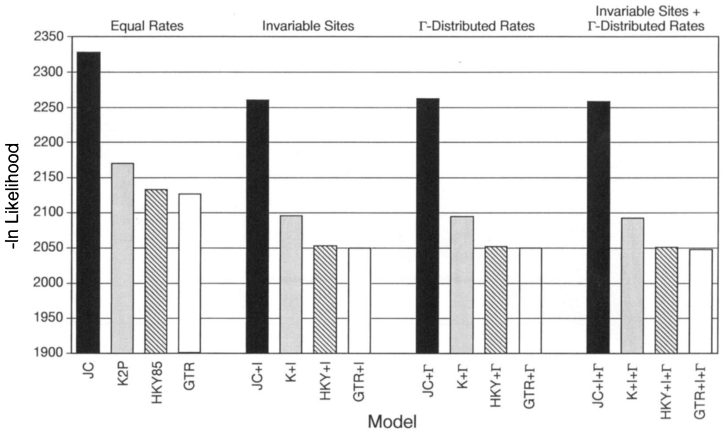
where $\ln L_0$ is the likelihood score under the more general model and $\ln L_1$ is the score of the restricted model.

The test statistic is asymptotically χ^2-distributed, with degrees of freedom equal to the difference in number of parameters between the models being compared.

This was the first approach used in model selection for phylogenetics, and the first two papers that actually implemented it are shown below.



Frati et al. (1997. J.Mol.Evol. 44:145)



Sullivan et al. (1997. Syst. Biol. 46:426)

We should point out that there are problems with using the χ^2-distributed with boundary values that are usually ignored.

Here's what I mean by this. The GTRer is a special case of GTR+I; p_{invar} is constrained to zero, which is one of the bounds of the values that p_{invar} can possibly assume.

This has the effect of bunching the distribution of the test statistic at its left end, so if we ignore the boundary value problem, we're going to risk rejecting an adequate simpler model too often.

This is dealt with by using a mixed 50/50 mixture of χ^2-distributions with zero and one d.f. as the null distribution.

Note that LRTs are restricted to pairwise comparisons. This means that we must traverse model space via a series of pairwise comparisons.

So how should we traverse model space? If we look at our cloverleaf diagram for the GTR+I+Γ family, two alternatives immediately become obvious. **Bottom up or top down.**

Either way we go, it's obvious that we need to make decisions about how we traverse this model space.

If we start at the bottom, with the JC model, we have to decide if we want to first relax the assumption of a single substitution type, the assumption of equal base frequencies, or the assumption of equal rates among sites.

In the first case, we would test JC vs. K2P with a χ^2 and 1 d.f.
In the second, we would test JC vs. F81 with a χ^2 and 3 d.f.
In the third, we would test JC vs. JC+I or perhaps JC+Γ, in either case with the mixed χ^2.

Say we decided on the first option and reject JC in favor of K2P. We now can relax the assumptions of K2P, by say allowing 3 substitution types, maybe allowing each transition to have its own relative rate parameter in the **R** matrix.

Usually, one stops adding parameters as soon as there's a step in the traversal of model space at which one can't reject the simpler model in favor of its more general companion.

This often results in not considering models that lie in a large portion of the model space, and it's common that if you take different pathways through model space, you'll end up with different models.

Joe gives an example of this on pages 327 & 328. A well-known example is Cunningham et al. (1998. Evolution, 52:978).

An alternative to the bottom-up approach is a top-down approach in which you start at the most general and parameter rich model you're considering, say the GTR+I+Γ model and try to simplify it.

You still have decisions to make regarding which aspect of the model you want to try to simplify first, and again, you can get to different models if you simplify differently.

One way around this is to look at the parameter estimates themselves, and let their values suggest how to proceed in simplifying.

Bayes Factors are Bayesian analogues of hLRT's, but they can be built into Bayesian MCMC.

IV. Novel Approaches

A. Performance Based Model Selection – Can we remove the exclusive reliance on fit?

We've developed a method that uses Decision Theory to incorporate estimates of branch length error along with a BIC in selecting models (Minin et al., 2003. Syst. Biol. 52:674).

Assume we are faced with a choice between, say, *A* or *B* and there are multiple possible outcomes of each choice.

There are *a* possible outcomes if choice *A* is made ($A_1, A_2,,,,A_a$).

And there are *b* possible outcomes if choice *B* is made ($B_1, B_2,...B_b$).

We need to quantify the cost associated with each potential outcome (i.e., $C_{A1}, C_{A2},...C_{Aa}$ and $C_{B1}, C_{B2},...C_{Bb}$).

Further, the approach requires that we determine the probabilities of each potential outcome, given each choice. That is, we must be able to calculate the following: $P(A_1|A)$, $P(A_2|A),...P(A_a|A)$ and $P(B_1|B), P(B_2|B),...P(B_b|B)$.

We then calculate the expected cost (i.e., the risk) associated with each choice.

$$R_A = \sum_{i=1}^{a} C_{A_i} P(A_i \mid A)$$

$$R_B = \sum_{i=1}^{b} C_{B_i} P(B_i \mid B)$$

We choose the model with the minimum risk.

We assume an unrooted phylogeny with k terminal nodes. Therefore, there will be $2k$-3 branches.

$\mathbf{B} = (B_1, B_2, \ldots, B_{2k-3})$ is the vector of branch lengths and $\hat{\mathbf{B}}_i$ is the vector of *estimated* branch lengths under the assumptions of model M_i.

If we have two models, M_i and M_j, the Euclidean distance between the branch length estimates is given by:

$$\left\| \hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j \right\|^2 = \sum_{l=1}^{2k-3} (\hat{\mathbf{B}}_{i,l} - \hat{\mathbf{B}}_{j,l})^2$$

and the risk of choosing model M_i is given by:

$$R_i = \sum_{j=1}^{m} \left\| \hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j \right\|^2 P(M_j \mid D)$$

Recall that the BIC_j is related to the posterior probability of model j. Thus,

$$R_i \approx \sum_{j=1}^{m} \left\| \hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j \right\|^2 \frac{e^{-BIC_j}}{\sum_{j=1}^{m} e^{-BIC_j}}$$

As detailed earlier, we can use an approximate (e.g., NJ) tree.

DT method incorporates fit (as measured by the likelihood score in calculation of the BIC), a penalty for over-parameterization, and expected branch-length error in selecting a model from among a set of candidates.

This method also compares all models simultaneously, so avoids the necessity to traverse model space with a series of pairwise comparisons.

In many instances the same model is chosen by this method as by other methods, but when they differ, the DT method chooses simpler models that nevertheless perform as well or better than models chosen by other methods.

B. Model Averaging

Since model selection represents a choice that is made with uncertainty, one view is that we should incorporate this uncertainty into our phylogeny estimation → Multimodel Inferences.

Although the ideal of model averaging in phylogenetics has been around for a while, there are just a few papers on it.

Posada & Buckley (2004. Syst. Biol. 53:793) reviewed model selection and recommend the use of AIC weighting in model averaging.

Here, the weights are based on the following:

$$\Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{min},$$

which is the difference in AIC score between M_i and the best AIC score in your set of candidates.

$\Delta_i \leq 2$ indicates substantial support for M_i.

$4 \leq \Delta_i \leq 10$ indicates weak support for M_i.

$\Delta_i \geq 10$ indicates no support for M_i.

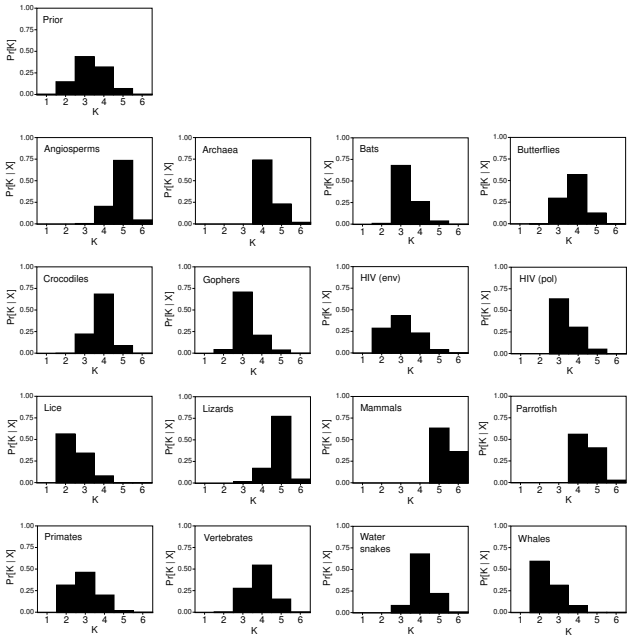
Inferences are made under each model with $\Delta_i \leq 2$ and the weighted average (of the tree or nodal support values, etc.) is derived.

Perhaps a better approach is to automate model averaging via reversible jump MCMC.

Reversible jump MCMC allows estimation of the models and accounting for model uncertainty directly and simultaneously to tree estimation.

Huelsenbeck et al. (2003. Mol. Biol. Evol. 21:1123) developed ModelJumper to do account for the 203 special cases of the GTR family, and Evans and Sullivan (2010. Mol. Biol. Evol., 28:343) have generalized the approach to permit evaluation of the equal base-frequency analogues as well (so all 403 special cases of GTR - actually both papers include a Γ-shape parameter as well, but ours also allows removal of gamma distributions, so includes 806 possible models).

Here, the frequency of each model in the posterior distribution permits an evaluation of $P(M_i \mid D)$. The figure below is from Huelsenbeck et al. (2003).



This is probably a better approach to model averaging than using the AIC weights, it's likely that the future of phylogenetics (at least Bayesian approaches) will use rjMCMC to integrate uncertainty in model choice into phylogeny estimation.

V. Does it matter how we select models?

So, we have this array of model-selection approaches that differ in their philosophy and formulation?

How different are:

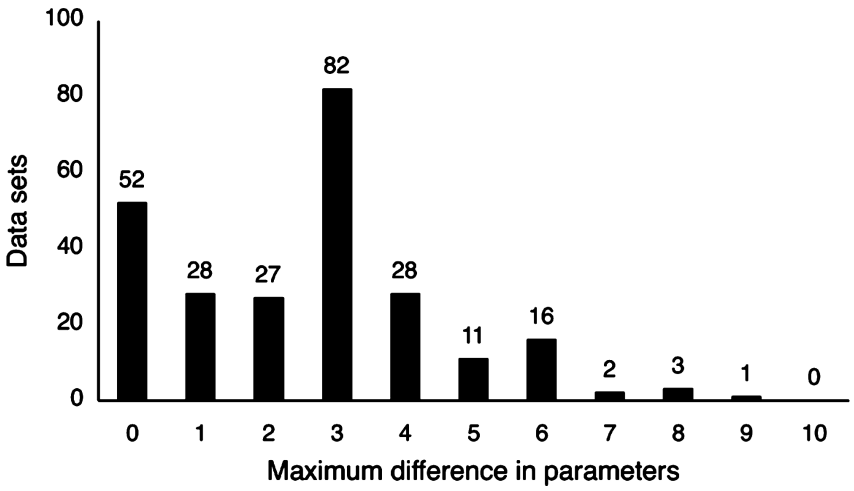
1) Models selected using the various approaches?
2) Inferences made using models selected with different approaches?

Jennifer Ripplinger (Ripplinger and Sullivan. 2008. Syst. Biol 57:76) addressed this in the first chapter of her dissertation by downloading 250 phylogenetic data sets from TreeBASE and selecting model using hLRT, AIC, BIC and DT.

All four picked same model in 51 data sets.
Two models were selected in 123 data sets.
Three were selected in 70 data sets.
All picked different models in 6 data sets.

Approach	avg # param
hLTR*	6.9 ± 2.2
AIC	8.4 ± 1.8
BIC	6.7 ± 1.7
DT	6.7 ± 1.4

The most commonly-used approach, AIC, tended to pick more complex and parameter rich models.

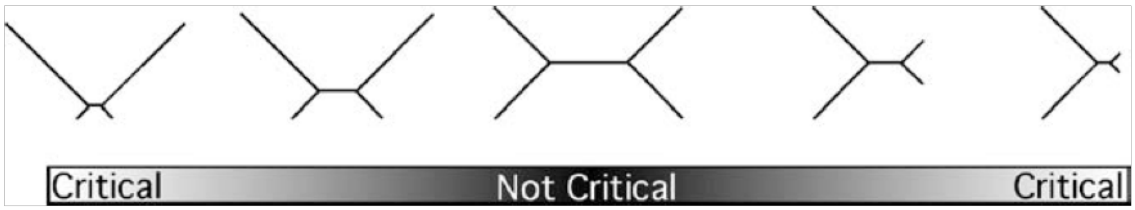


Model selected with different approaches often were quite different from each other with>half the cases differing by 3 or more parameters.

Furthermore, when models differed, ML trees differed ~50 % of time.

However, those topological differences were usually restricted to poorly supported nodes, so model selection is critical only under some conditions.

Sullivan and Joyce (2005) concluded that these conditions can arise because of extreme branch-length disparity.



The conclusions of a recent paper (Abadi et al. 2019. Nature Communications) are due to the fact that they set up their simulations in the middle of this continuum.

V. A Diversion into No Common Mechanism

Since Nick Goldman first published a formal description of parsimony in terms of a likelihood function (Goldman. 1990. Syst. Zool. 39:345), we've had a single mathematical framework for interpreting the two approaches.

Specifically, he showed that if all 2*n*-3 branch lengths are equal, ML under a JC model is equivalent to a parsimony model (i.e., under these conditions, parsimony is an ML estimator).

Of course, however, there's no reason to expect that all branches in a phylogeny have the same lengths, but the formulation of a likelihood model for parsimony does allow us to compare the two approaches directly.

Tuffley & Steel (1997. Bull. Math. Biol. 59:581) generalized this to a No Common Mechanism model (NCM). Here, they use a JC substitution process, but apply a separate JC to each branch for each site. Further, they indicated that this is also equivalent to a parsimony model.

This can be seen in this 2008 paper by Huelsenbeck et al. (Syst. Biol. 57:306).

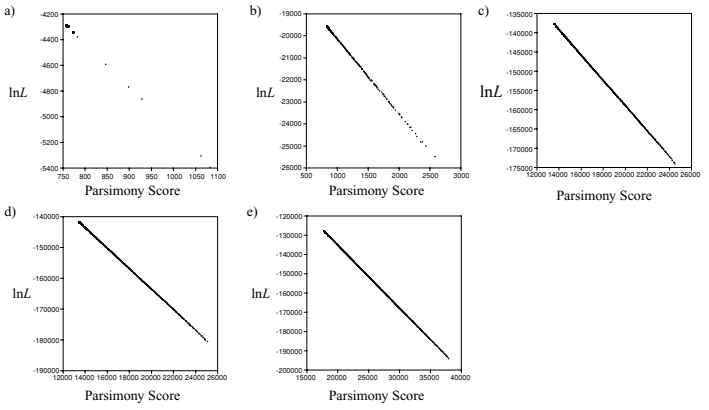
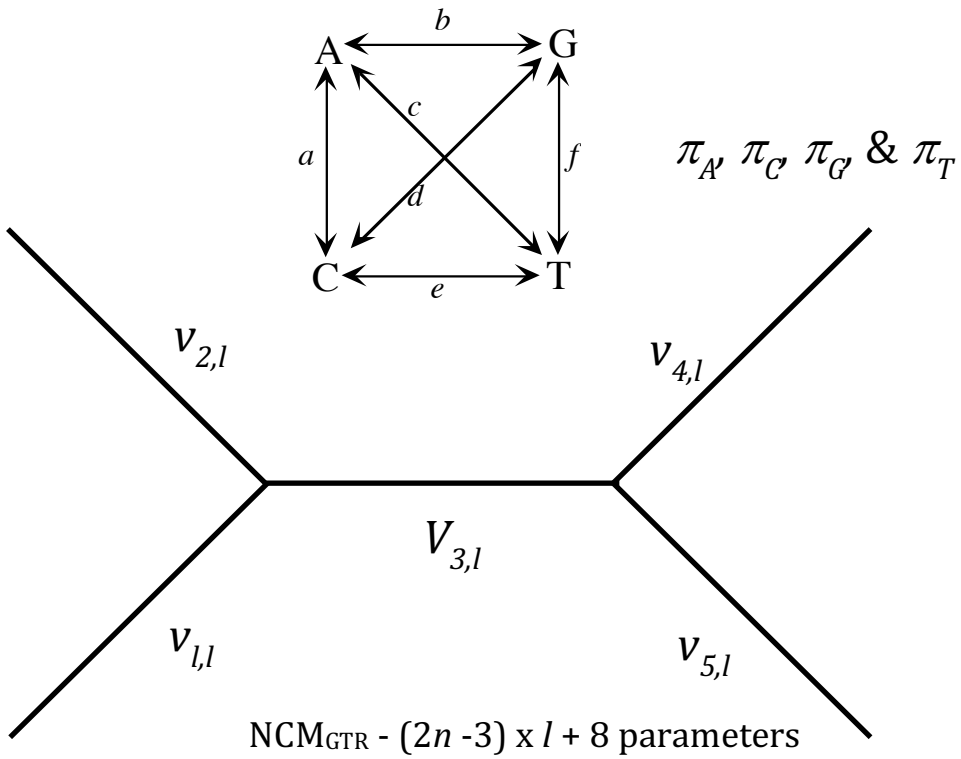


FIGURE 5. The relationship between the parsimony score and the log likelihood of a tree under the no-common-mechanism model of Tuffley and Steel (1997) for (a) the vertebrate β-globin alignment (Yang et al., 2000); (b) the *Astragalus* ITS alignment (Sanderson and Wojciechowski, 2000); (c) the Angiosperm *rbcL* and (d) *atpB* alignments (Savolainen et al., 2000); and (e) the alignment of *rbcL* gene sequences for green plants (Chase et al., 1993). Each plot contains the 20,000 trees sampled using the stochastic NNI tree proposal mechanism.

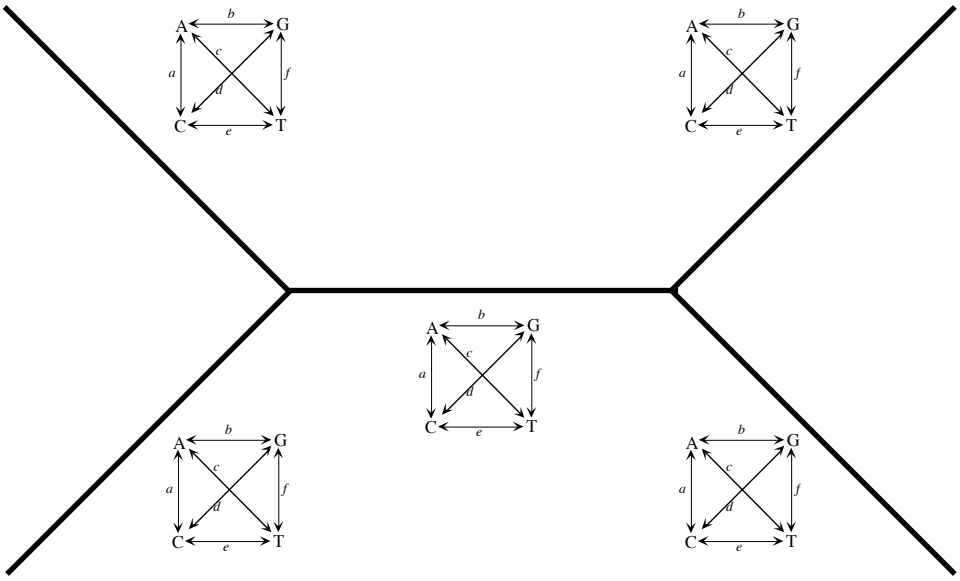
Thus, for the Tuffley-Steel version of NCM, we'll have (2*n*-3) x *l* parameters for an analysis with *n* taxa and *l* nucleotides.

Huelsenbeck et al. (2011. Syst. Biol. 60:1) developed two elaborations of this involving the GTR model.

> First, they allowed a single **Q**-matrix to allow for non-equal base frequencies and six substitution types, with **Q** being applied across all sites and across all branches, but with a variable length for each branch and site combination.



Second, they took the rational of NCM to its logical conclusion and allowed a separate GTR for each branch and site combinations.



NCM_{GTRC} - $(2n\text{-}3)$ x l x 8

A couple papers have evaluated the NCM models from a model-selection perspective.

1) Holder et al. (2010. Syst. Biol. 59:477) demonstrated that the AIC will never favor the original (JC) variant of the NCM model over any of the common mechanism models.

2) Hulesenbeck et al. (2011. Syst. Biol.) evaluated the marginal likelihoods of several GTR+Γ special cases relative to the various NCM variants.