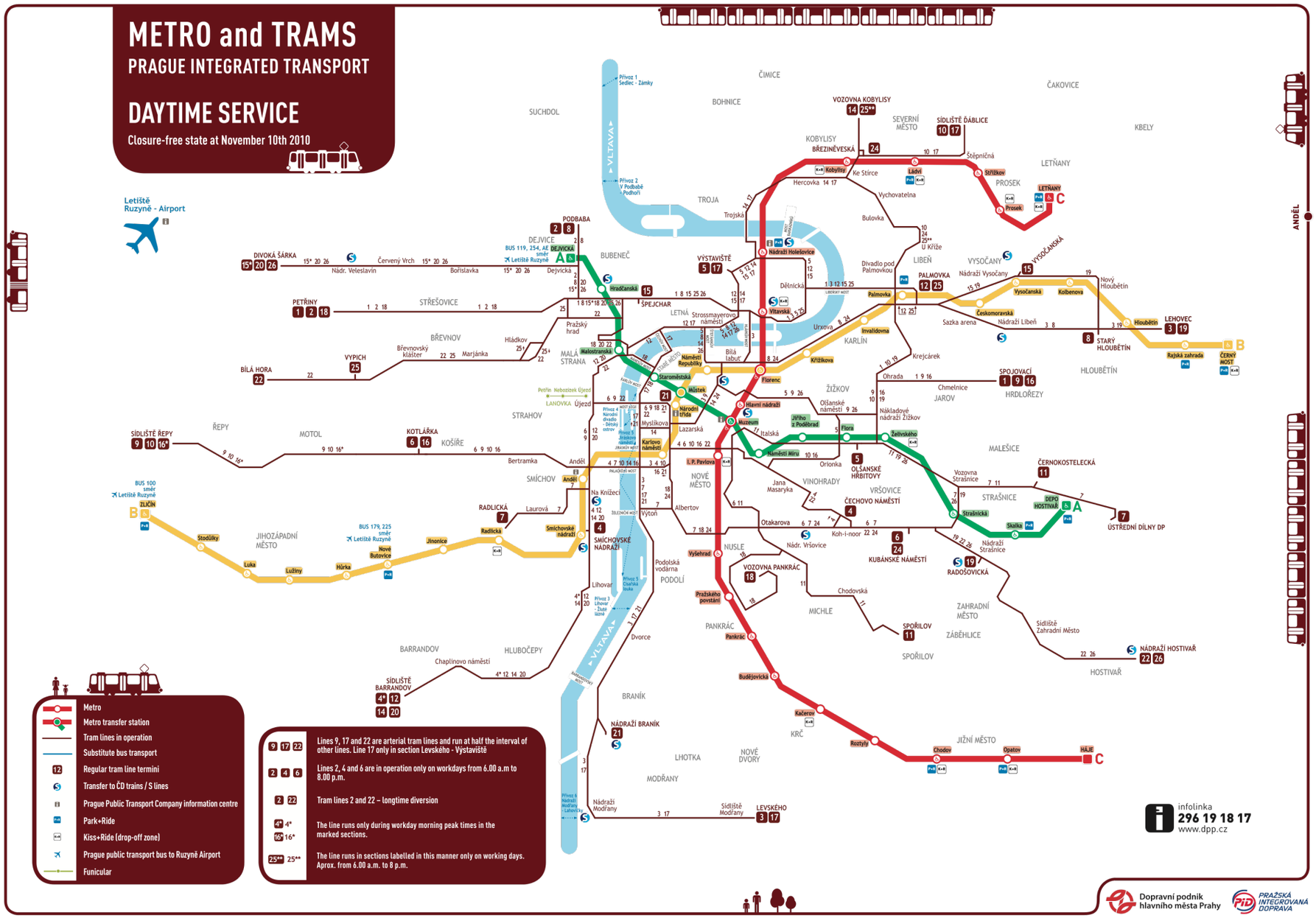
Model Selection in Phylogenetics

David L. Swofford

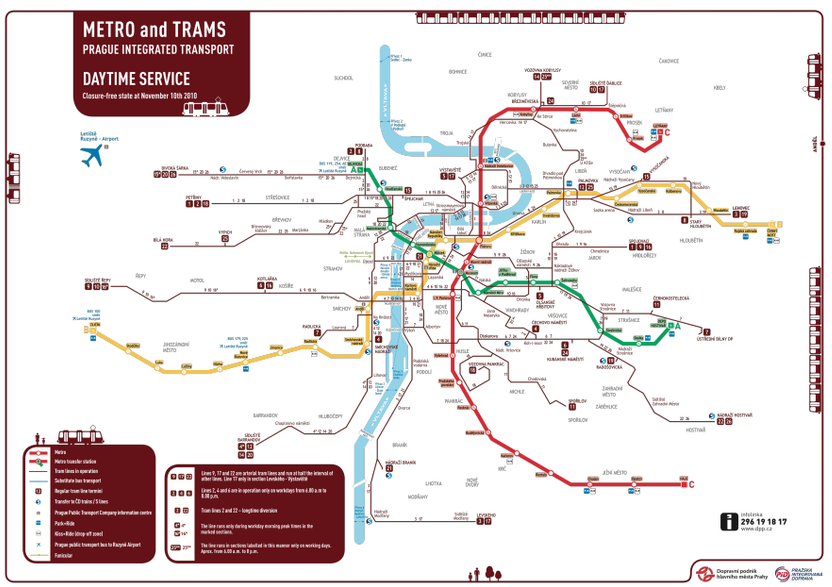
Duke University

A model of the Prague Metro



A less complex model of the Prague Metro

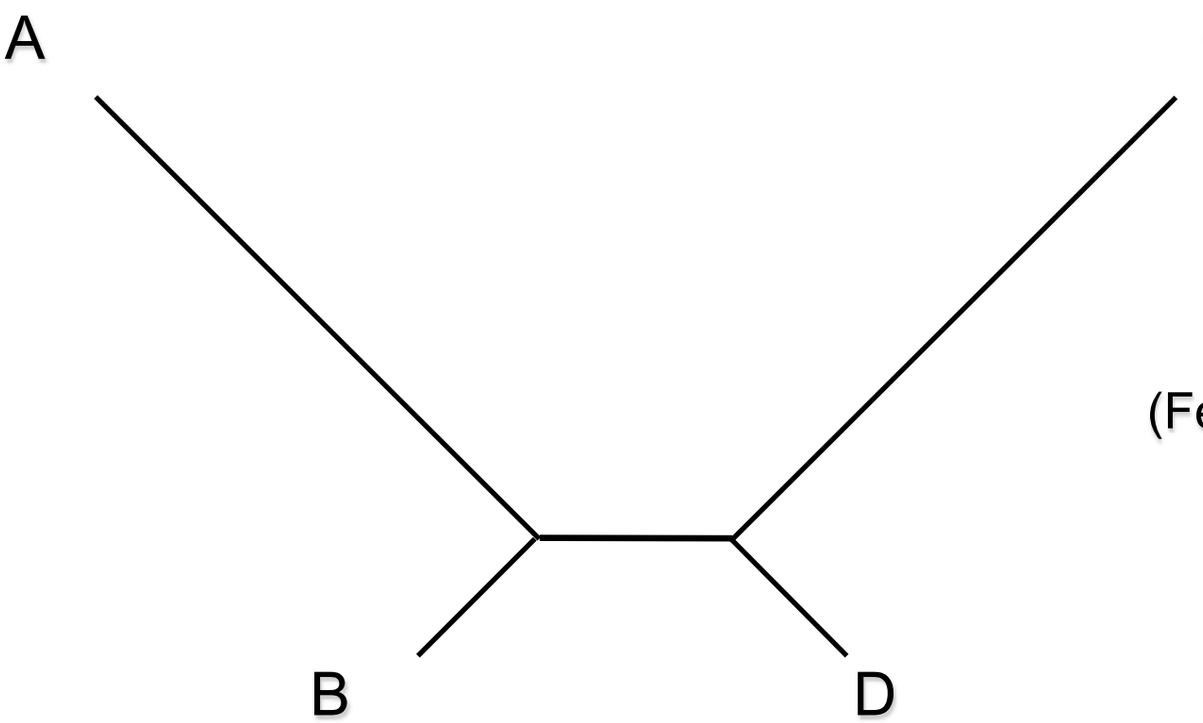
Which model is most useful?



When do models matter in phylogenetics?

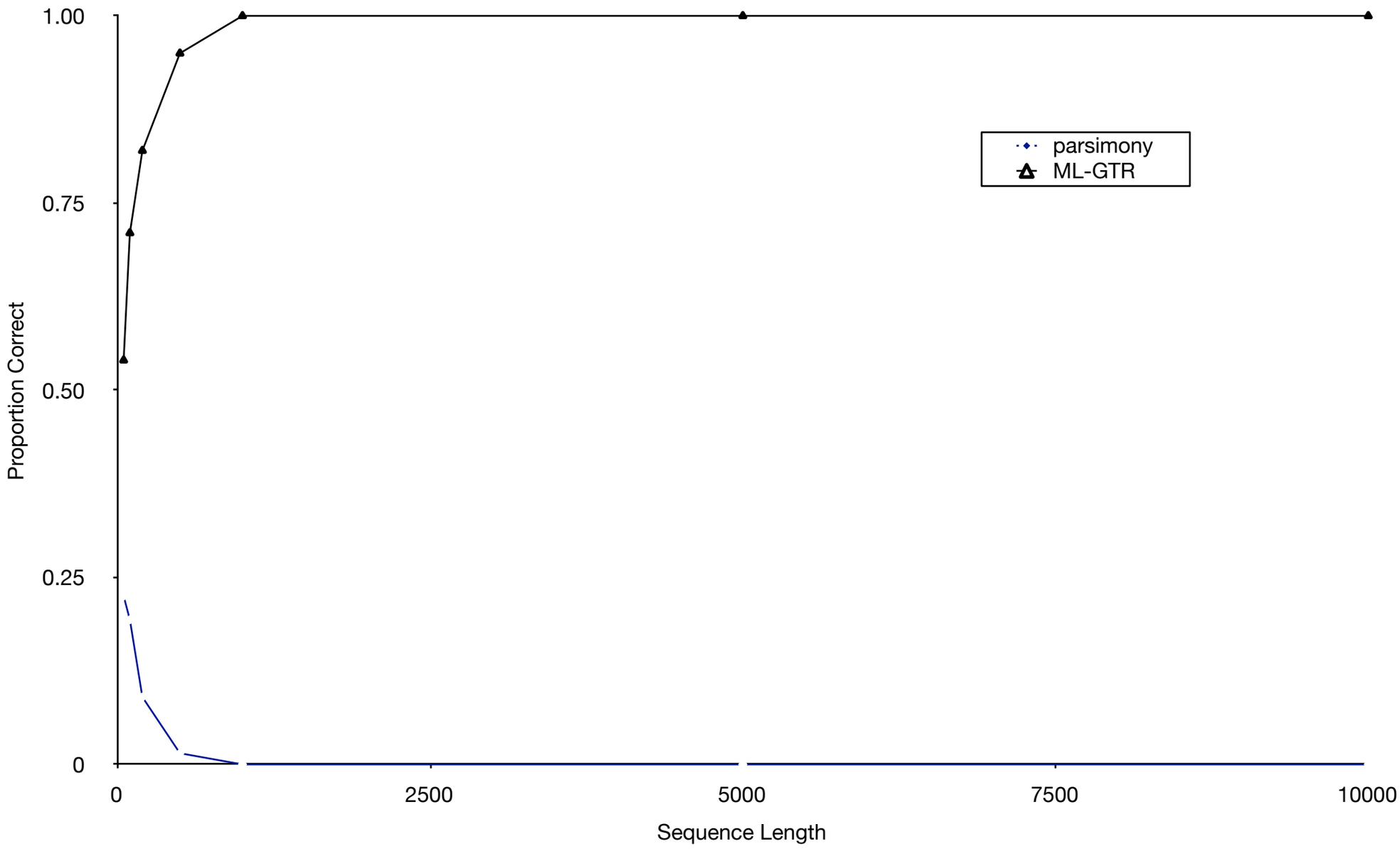
- Model-based methods including ML and Bayesian inference (typically) make a *consistent* estimate of the phylogeny (estimate converges to true tree as number of sites increases toward infinity)

... even when you're in the "Felsenstein Zone"



(Felsenstein, 1978)

In the Felsenstein Zone



Simulation model = GTR

Why do models matter? (continued)

Parsimony is inconsistent in the Felsenstein zone (and other scenarios)

Likelihood is consistent in any "zone" (when certain requirements are met)

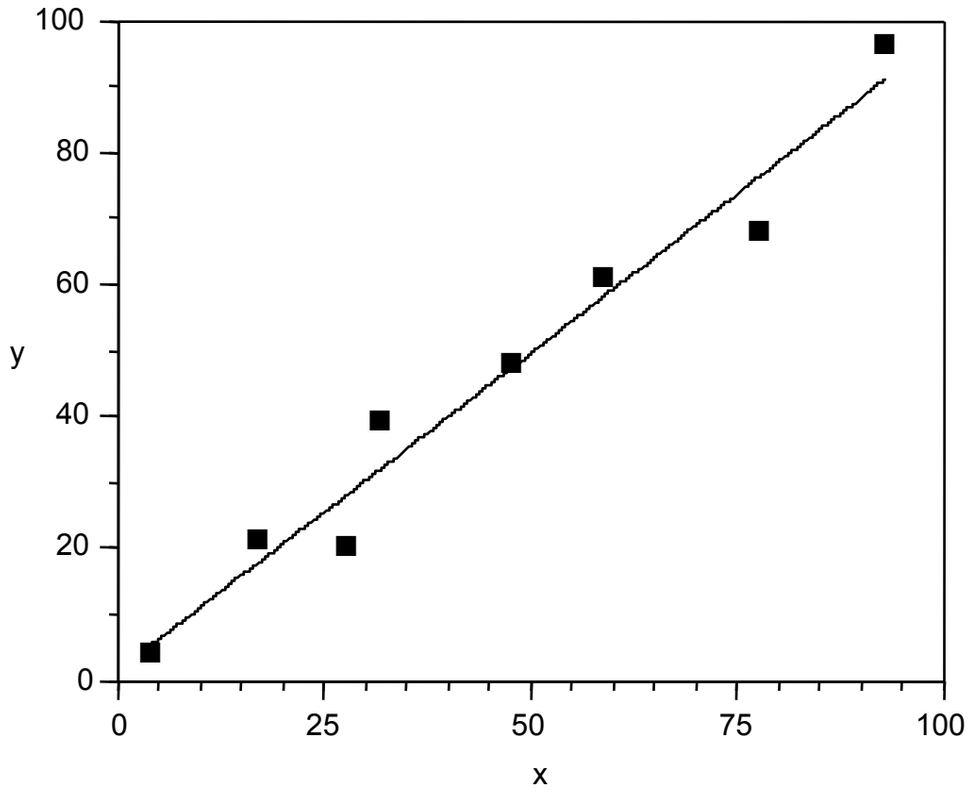
But this guarantee requires that the model be specified correctly! Likelihood can also be inconsistent if the model is oversimplified

Real data always evolve according to processes more complex than any computationally feasible model would permit, so we have to choose "good" rather than "correct" models

What is a “good” model?

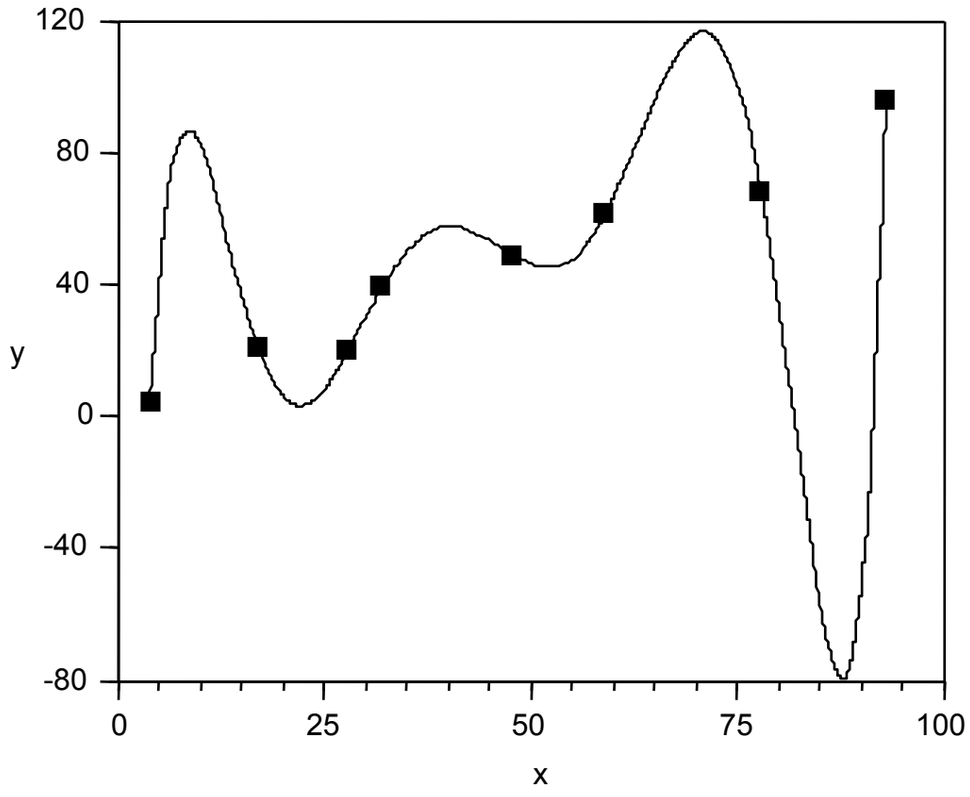
A model that appropriately balances fit of the data with simplicity (parsimony, in a different sense)

i.e., if a simpler model fits the data almost as well as a more complex model, prefer the simpler one



$$y = 1.30 + 0.965x$$

$$(r^2 = 0.963)$$



$$y = -330 + 134x - 15.5x^2 + 0.816x^3$$
$$- 0.0225x^4 + 0.000335x^5$$
$$- 0.00000255x^6 + 0.00000000777x^7$$
$$(r^2 = 1.000)$$

“The Principle of Parsimony” in the world of statistics

Burnham and Anderson (1998): Model Selection and Inference

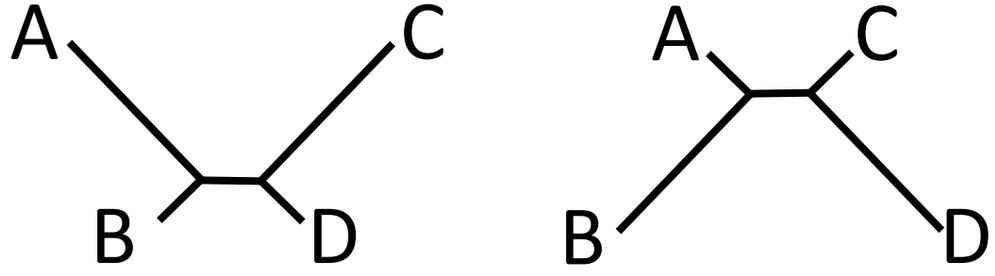
Parsimony lies between the evils of underfitting and overfitting. The concept of parsimony has a long history in in the sciences. Often this has been expressed as “Occam’s razor”—shave away all that is not necessary. Parsimony in statistics represents a tradeoff between bias and variance as a function of the dimension of the model. A good model is a balance between under- and over-fitting.

Why models don't have to be perfect

Assertion: In most situations, phylogenetic inference is relatively robust to model misspecification, as long as critical factors influencing sequence evolution are accommodated

Caveat: There are some kinds of model misspecification that are very difficult to overcome (e.g., "heterotachy")

E.g.:

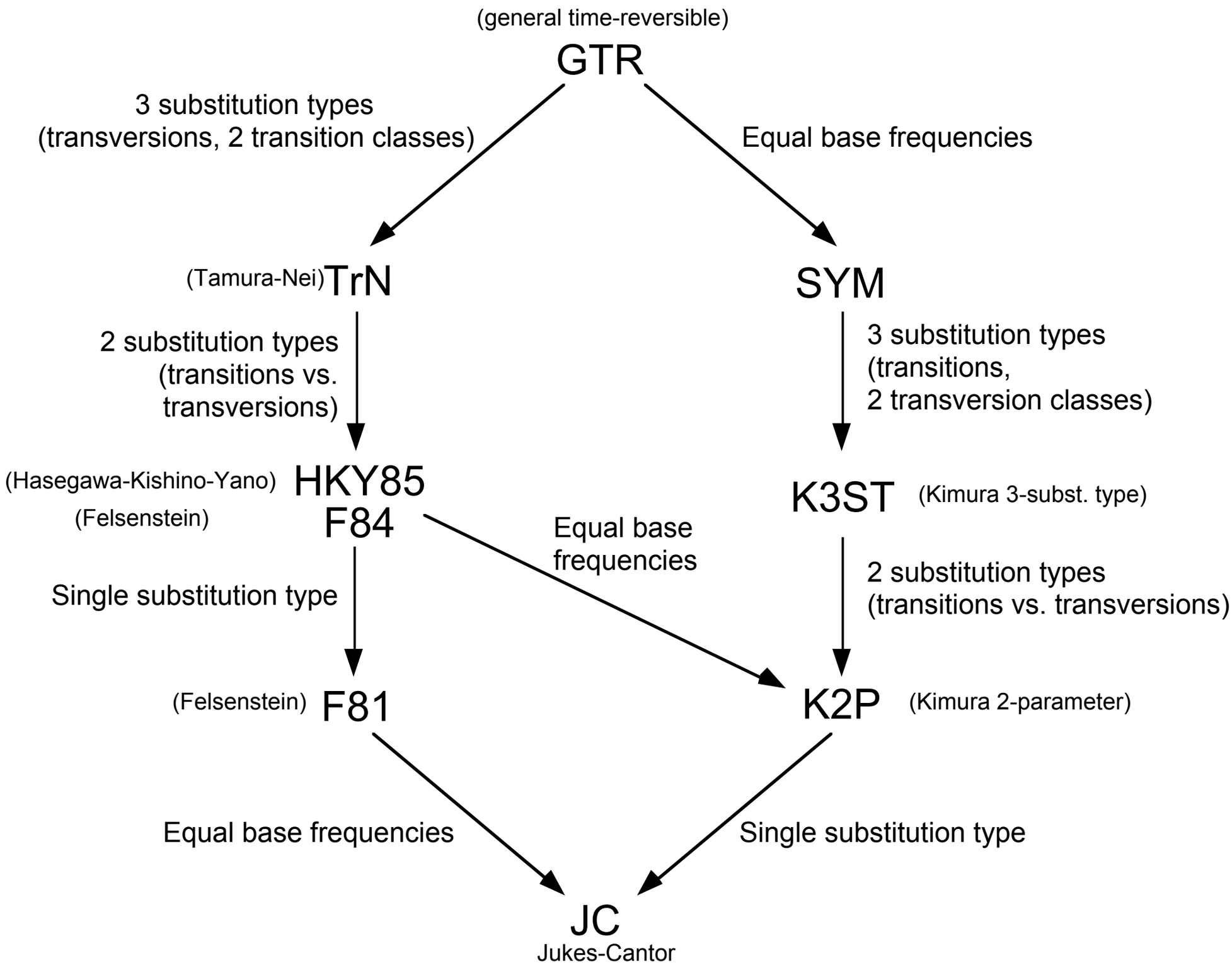


Half of sites

Other half

Likelihood can be consistent in Felsenstein zone, but will be inconsistent if a single set of branch lengths are assumed when there are actually two sets of branch lengths (Chang 1996)

GTR Family of Reversible DNA Substitution Models



Modeling among-site rate heterogeneity

equal rates? →

```
Lemur AAGCTTCATAG TTGCATCATCCA …TTACATCATCCA
Homo  AAGCTTCACCG TTGCATCATCCA …TTACATCCTCAT
Pan   AAGCTTCACCG TTACGCCATCCA …TTACATCCTCAT
Goril AAGCTTCACCG TTACGCCATCCA …CCCACGGACTTA
Pongo AAGCTTCACCG TTACGCCATCCT …GCAACCACCCTC
Hylo  AAGCTTTACAG TTACATTATCCG …TGCAACCGTCCT
Maca  AAGCTTTTCCG TTACATTATCCG …CGCAACCATCCT
```

- Proportion of invariable sites

  Some sites extremely unlikely to change due to strong functional or structural constraint (Hasegawa et al., 1985)

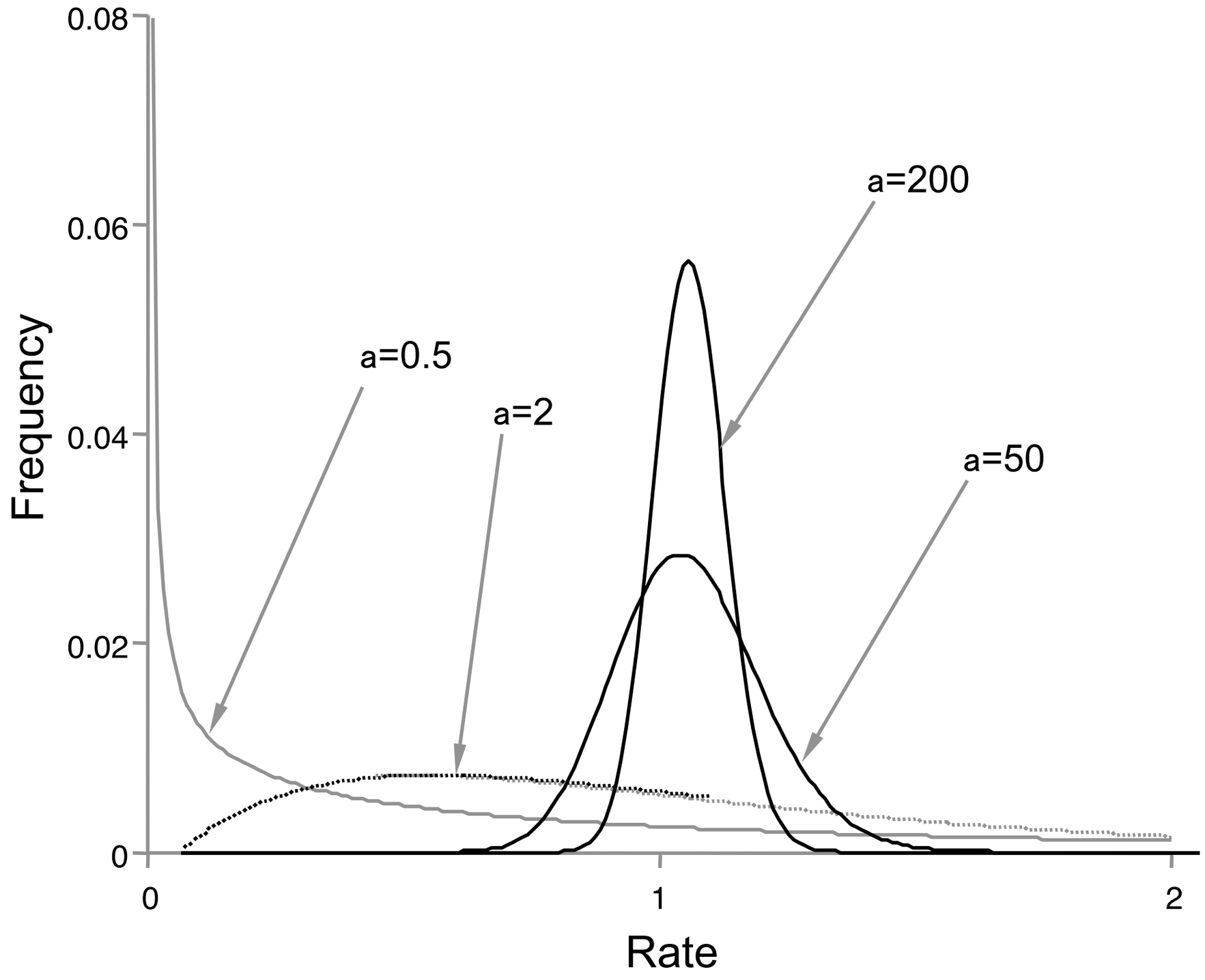
- Gamma-distributed rates

  Rate variation assumed to follow a gamma distribution with shape parameter α

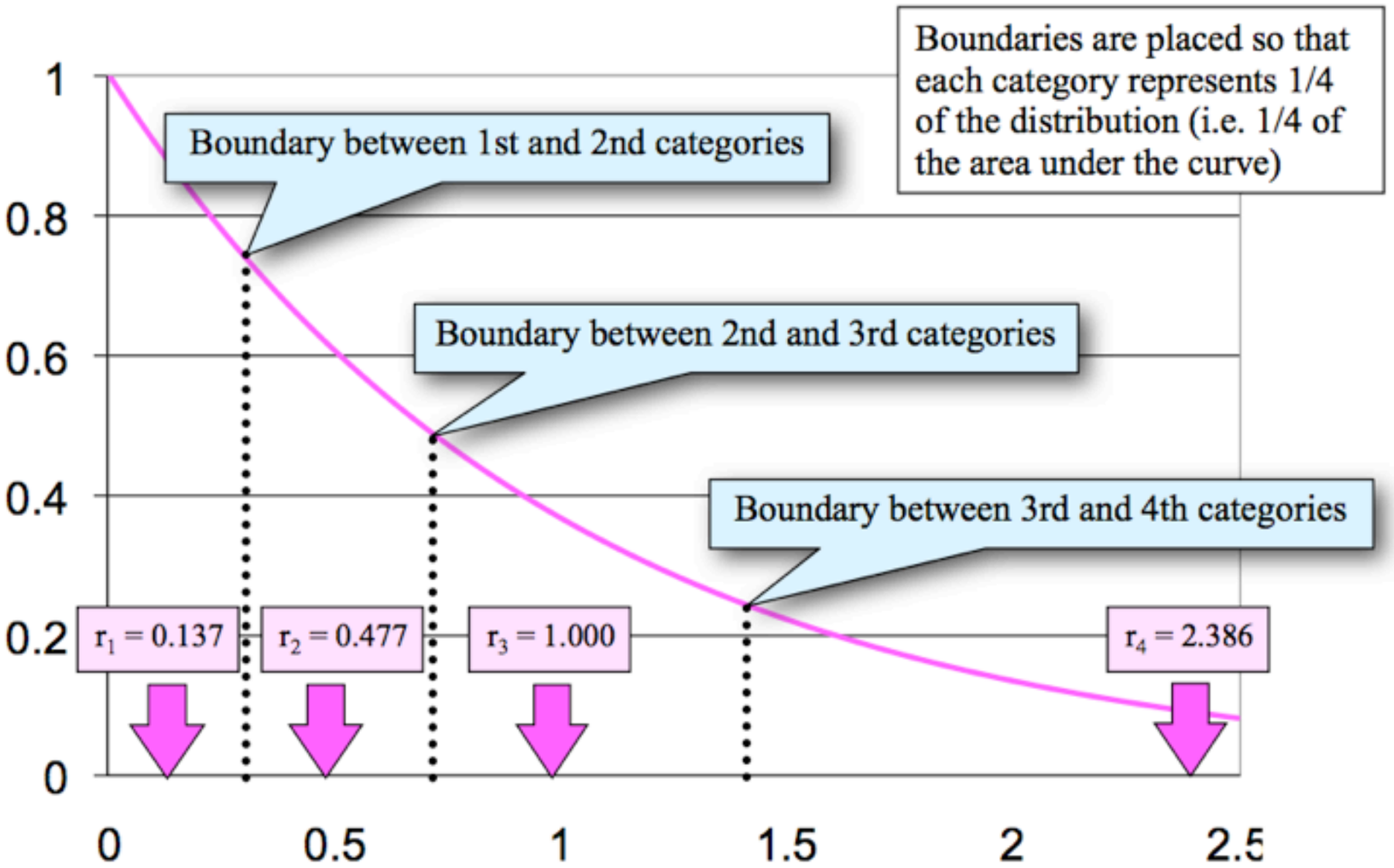
- Site-specific rates (another way to model ASRV)

  Different relative rates assumed for pre-assigned subsets of sites

Modeling ASRV with a gamma distribution ("+G")



For computational reasons we "discretize" the gamma distribution

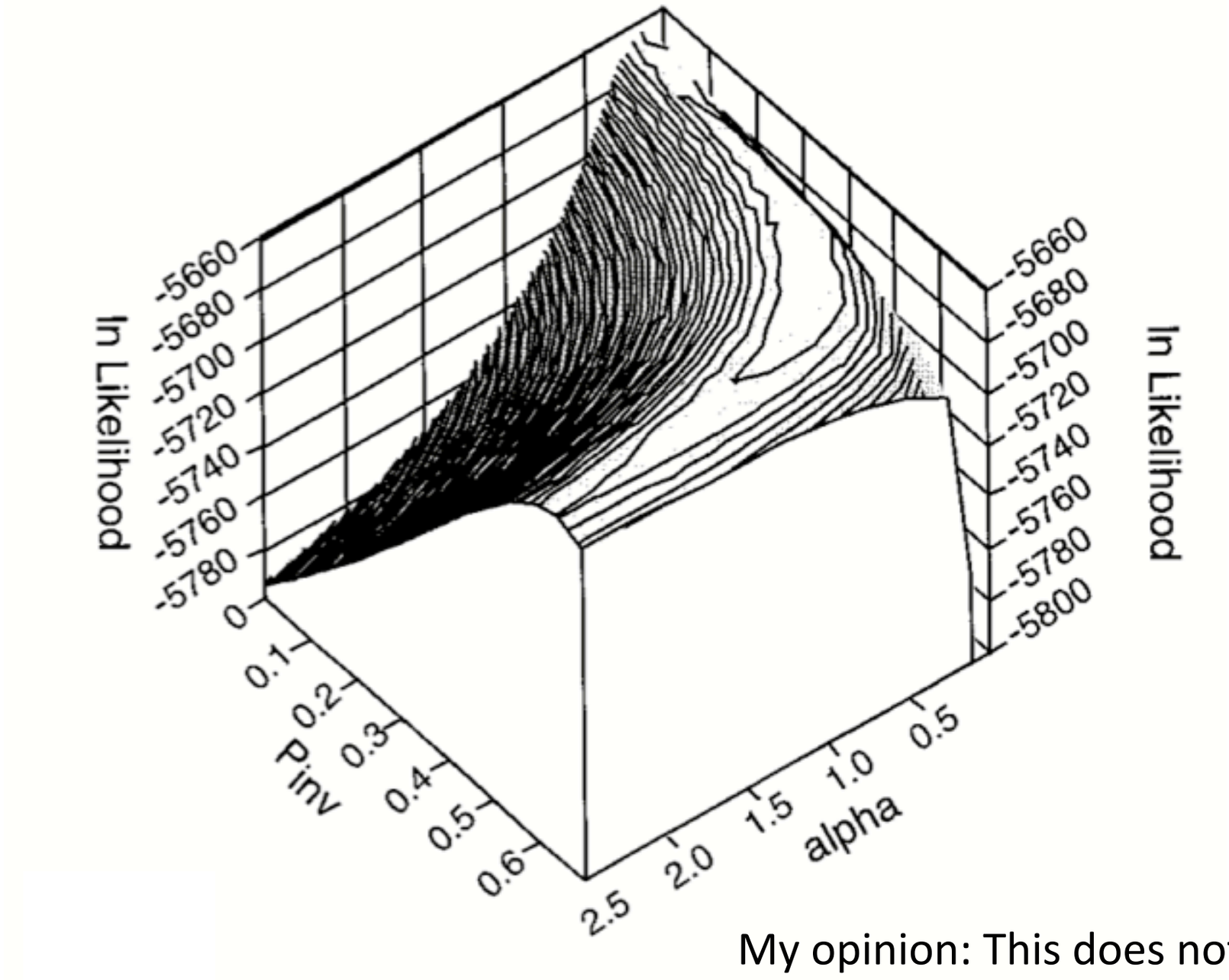


Slide from Paul Lewis's lecture

Can optionally also include an invariable sites category $r_0 = 0$ ("+I+G")

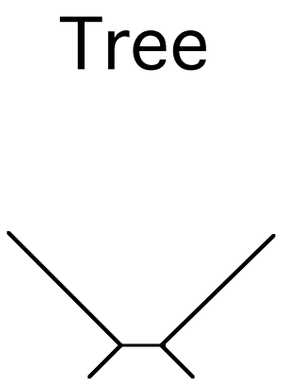
An aside on "+I+G" models

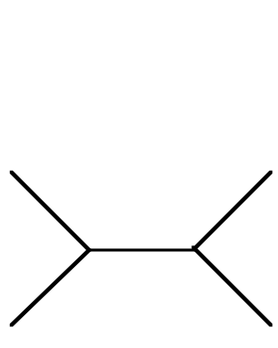
Models containing both gamma-distributed rates and invariable sites can be problematic due to the correlation of α and p_{inv}

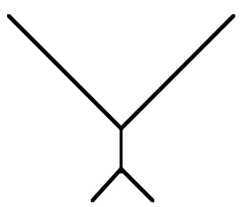


My opinion: This does not invalidate the use of the model, it just means that caution must be used in interpreting the parameter values

Performance of ML when its model is violated





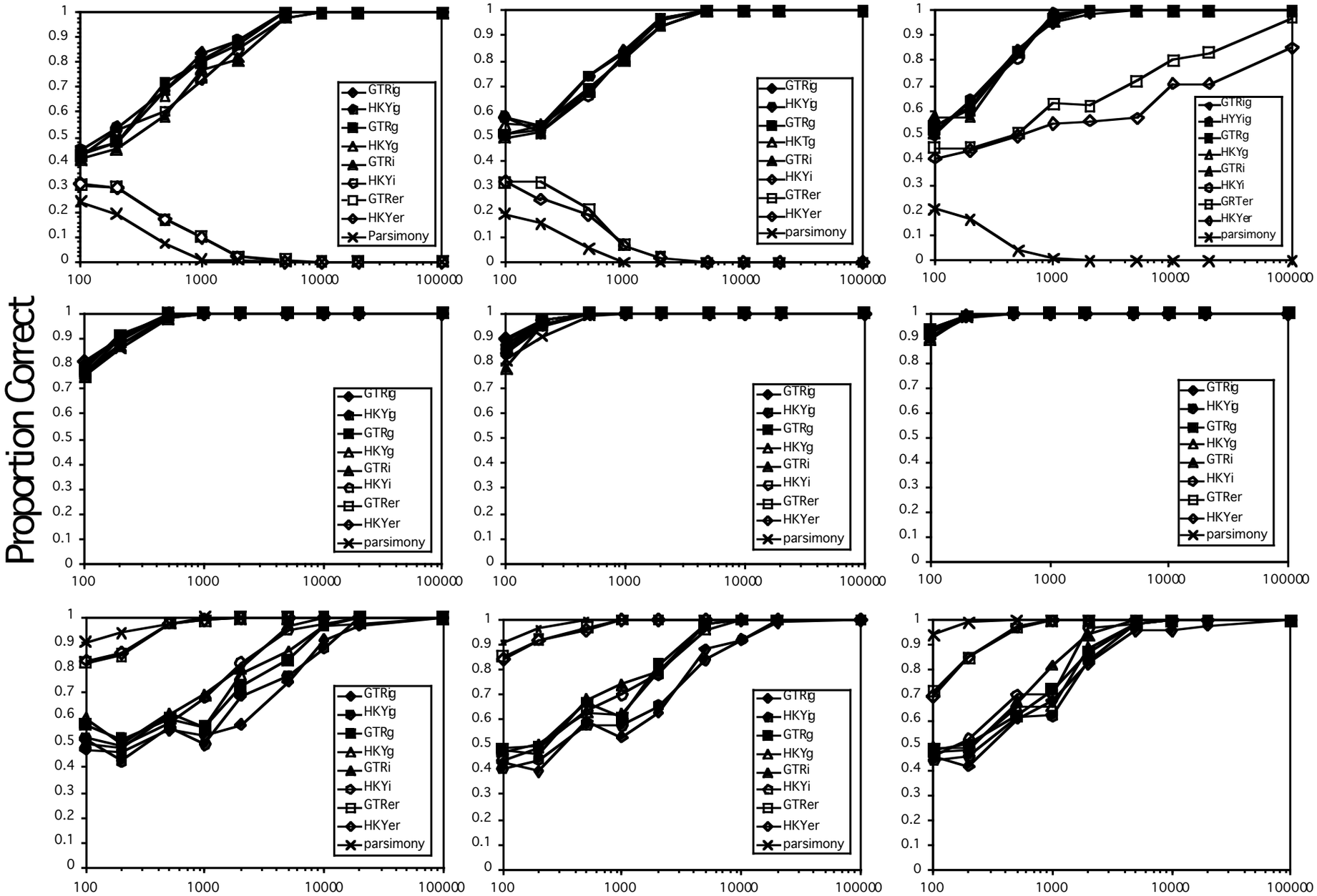


Tree

α = 0.5, pinv=0.5

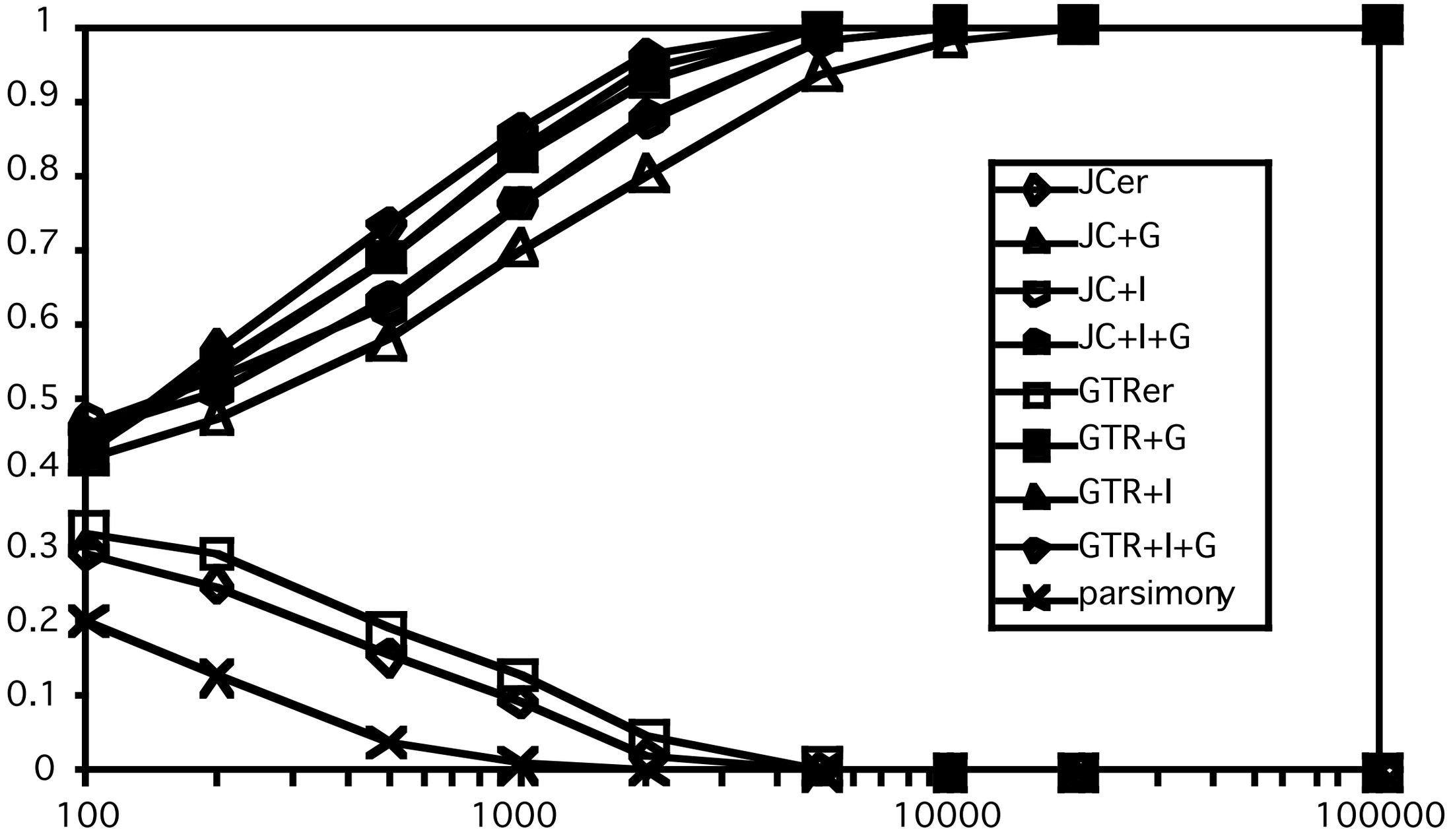
α = 1.0, pinv=0.5

α = 1.0, pinv=0.2

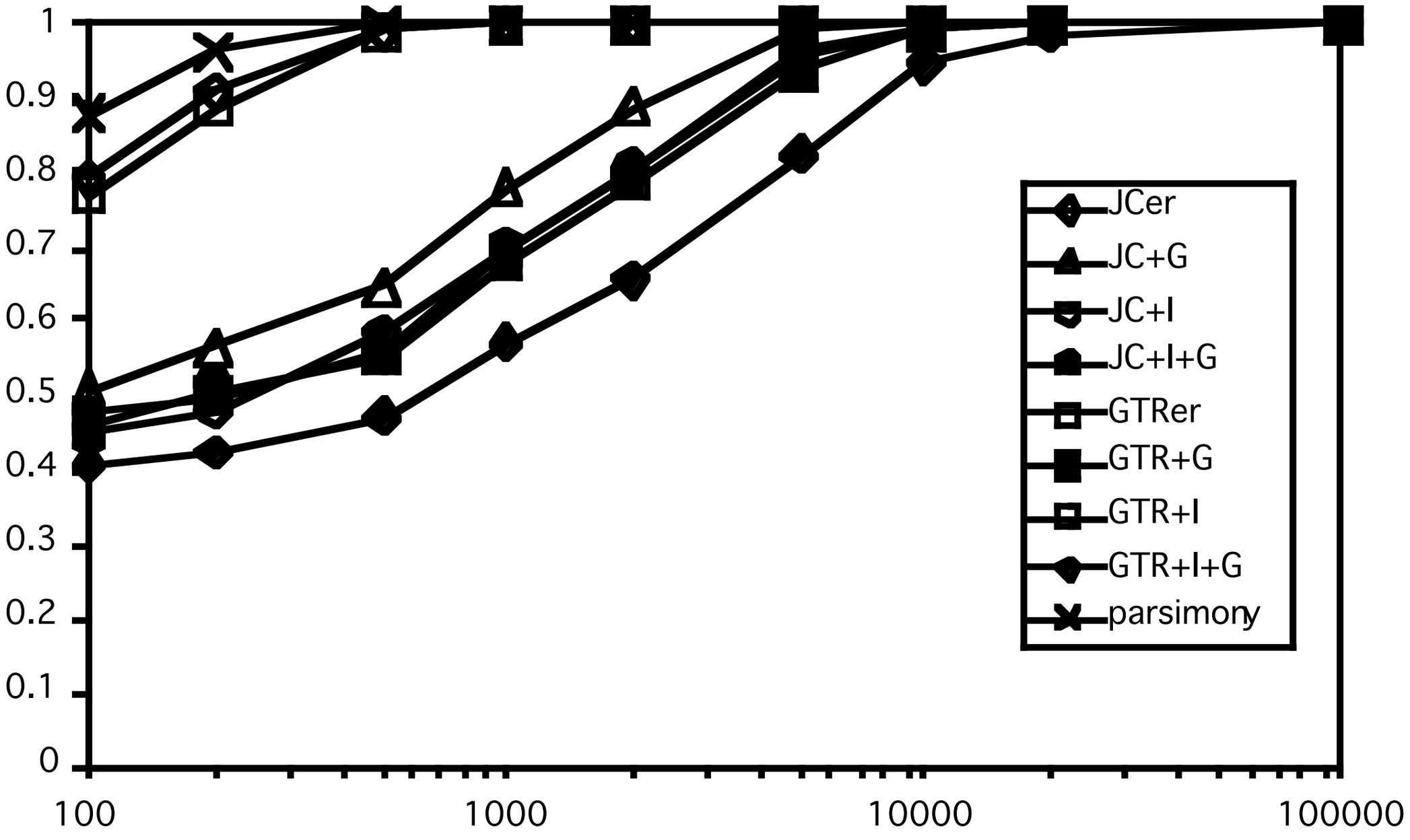


Sequence Length

“Moderate” rate variation
Felsenstein zone



“Moderate” rate variation inverse-Felsenstein zone



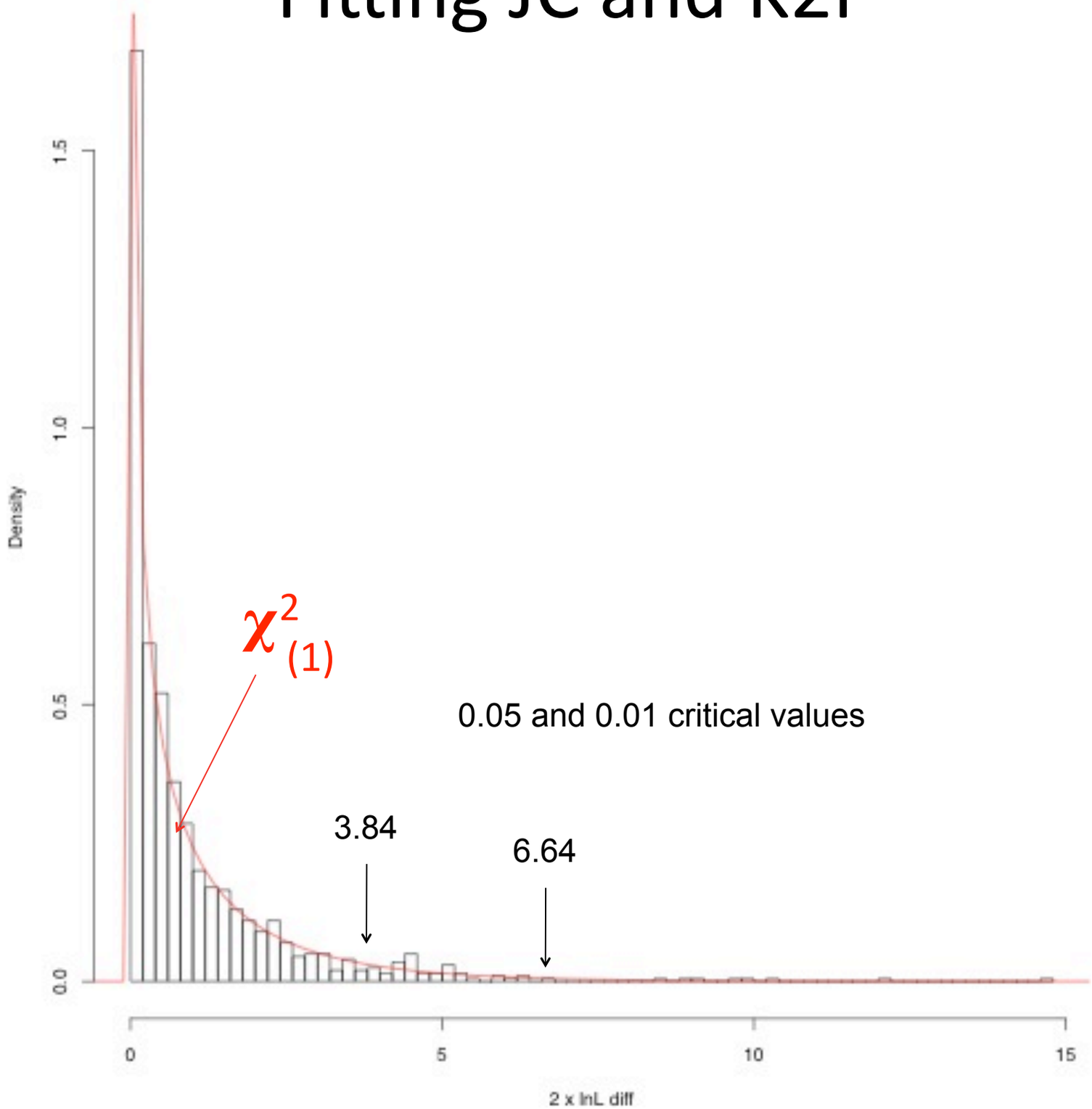
Model selection criteria

Likelihood ratio tests

Test statistic: $\delta = -2\left(\ln\ L_0 - \ln\ L_1\right)$

If model L_0 is nested within model L_1, δ is (asymptotically) distributed as X^2 with degrees-of-freedom equal to difference in number of free parameters

Simulation under JC
Fitting JC and K2P



Model selection criteria

Akaike Information Criterion (AIC):

$$AIC_i = -2 \ln L_i + 2K$$

where K is the number of free parameters estimated

$$AICc_i = AIC_i + \frac{2K(K+1)}{N-K-1}$$ ("corrected")

Model selection criteria

Bayesian Information Criterion (AIC):

$$BIC_i = -2 \ln L_i + K \ln N$$

where *N* is the "sample size" (typically number of sites)

Note than ln N exceeds 2 when K = 8, so BIC typically penalizes model complexity much more heavily

Partitioned Models

Up to now, we have been talking about homogeneous (unpartitioned) models, but many authors have emphasized the importance of modeling heterogeneity among genes or other subsets of the data appropriately

- Buckley, T. R., Arensburger, P., Simon, C., & Chambers, G. K. (2002). Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. *Systematic Biology*, *51*(1), 4-18.
- Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574.
- Suchard, M., Kitchen, C. M. R., Sinsheimer, J. S., & Weiss, R. E. (2003). Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology*, *52*(5), 649–664.
- Pagel, M., & Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, *53*(4), 571–581.
- Brandley, M. C., Schmitz, A., & Reeder, T. W. (2005). Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Systematic Biology*, *54*(3), 373–390.

... and many more recent papers

> "...data partitioning is more an art than a science, and it should rely on our knowledge of the biological system..."
>
> Yang and Rannala (2012; *Nature Rev. Genet*. 13:303-314)

Ways to partition

- By gene
- By codon
- By gene/codon combination
- Stems vs. loops (probably not advisable—e.g., Simon et al., 2006)
- Coding vs. noncoding

Naive partitioning

- Run ModelTest/JModelTest; estimate a model (from the GTR+I+G family) separately for each gene/subset
- Perform an ML/Bayesian analysis, assigning the chosen models to each

Too many parameters! 1-10 parameters for each gene; amount of data available to estimate each parameter does not increase

Over-Partitioning

Consider the following (contrived) example:

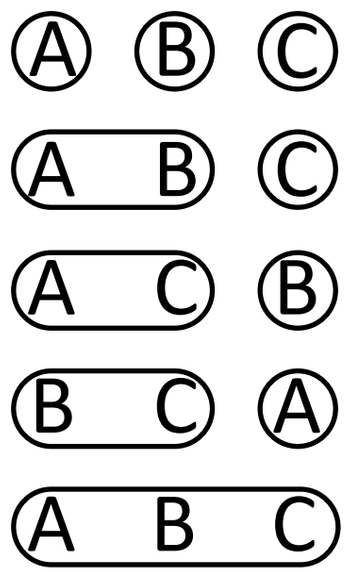
- Gene A: HKY+G, π = (0.26, 0.24, 0.23, 0.27), κ=1.1, α=3.0
- Gene B: GTR, π = (0.25,0.24,0.25,0.26), (a,b,c,d,e)=(1.1, 1.2, 0.9, 1.1, 0.95)
- Gene C: JC+I (pinv=0.05)

These are all GTR models that are not far from the Jukes-Cantor model, but they all have different “names”

Better to estimate one GTR model (even with 5+3+1+1=10 parameters, estimated from all data) than 3 separate models with 2+5+1=8 parameters (but only one gene’s worth of data for each model)

How to find optimal partitionings?

Consider a data set with 3 genes, A, B, and C:



For each partitioning scheme, evaluate some set of models from the GTR+I+G (e.g., 56 models) according to AIC or BIC

Choose a combination of partitioning scheme and model for subsequent partitioned-model analyses

Rob Lanfear's **PartitionFinder** (http://www.robertlanfear.com/partitionfinder/) automates this process; method now also available in PAUP* test versions

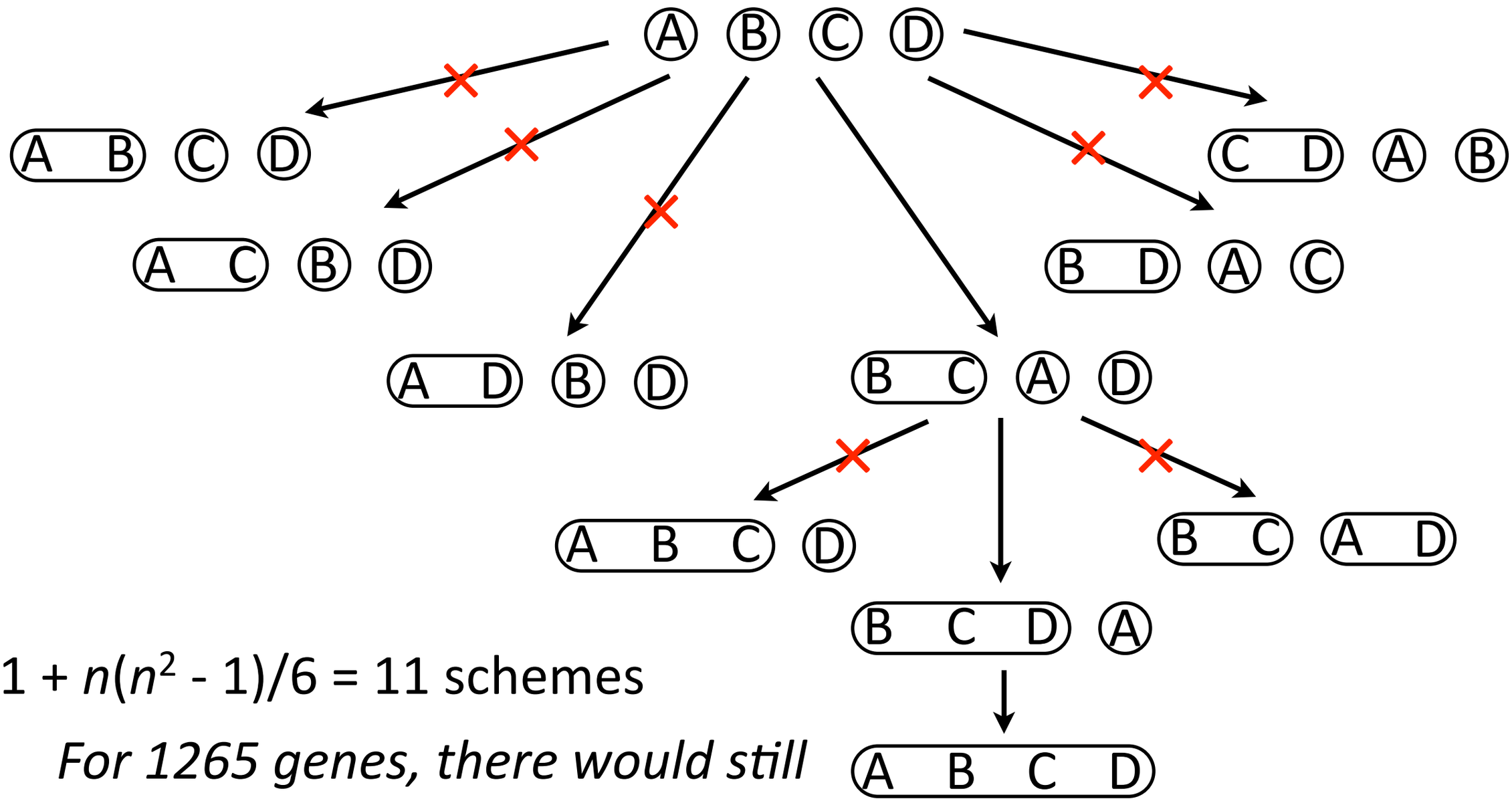
How many partitionings?

In general, the number of partitionings on n subsets is a “Bell number”

N	Bell number
2	2
3	5
4	52
5	203
6	877
7	4140
12	4×10^6
60	9.8×10^{59}

Clearly, there are too many partitioning schemes to evaluate them all for more than a few subsets.

Greedy algorithm when there are too many



$1 + n(n^2 - 1)/6 = 11$ schemes

For 1265 genes, there would still be 337,380,561 schemes to evaluate!

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, *29*(6), 1695–1701

How to partition thousands of genes (or other subsets)?

Cluster analysis

- Li, Lu, and Orti (2008)

  Estimate unlinked model parameters using a shared model; similar subsets will have similar parameter estimates and will cluster together.

  Problem? Similar models (in the sense of predicting similar site pattern frequencies), can have different parameter MLEs. Also must use same model for all subsets.

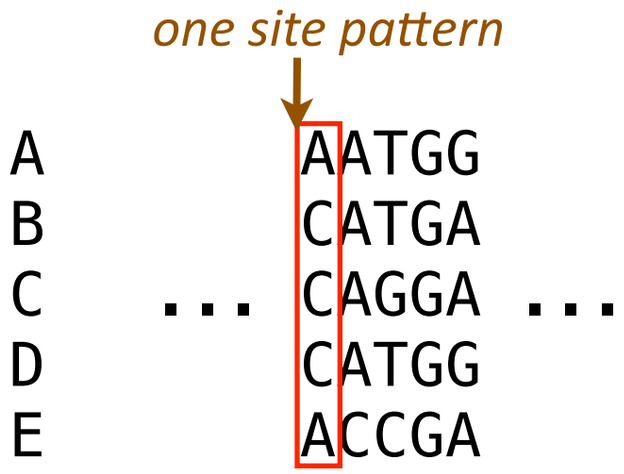
- Lanfear et al. (most recent PartitionFinder)

  Compute single-site likelihood values; cluster sites that have similar site likelihoods into larger subsets.

  Problem? Site likelihood is more determined by rate than anything else. Evolutionary rates will have more to do with the clusterings than differences in substitution pattern, state frequencies, etc.

A possible solution?

Cluster based on expected site-pattern frequencies



- Estimate an ML model for each subset
- Compute expected frequency of each possible site pattern according to chosen model

  In general, too many to use them all (4^T for DNA). Koch and Holder (2012) have an algorithm for calculating the probability that a site will be a member of a particular class of character patterns (number of parsimony steps plus nucleotides present); use this to objectively pool sites.
- Cluster together those subsets that have similar site-pattern-frequency spectra.

  Site pattern frequencies constitute a discrete probability distribution, so natural dissimilarity measure is the **Kullback-Leibler divergence**.

Clustering on Kullback-Leibler divergences

Can’t use simple K-means algorithm (KL divergences are non-symmetric and violate the triangle inequality)

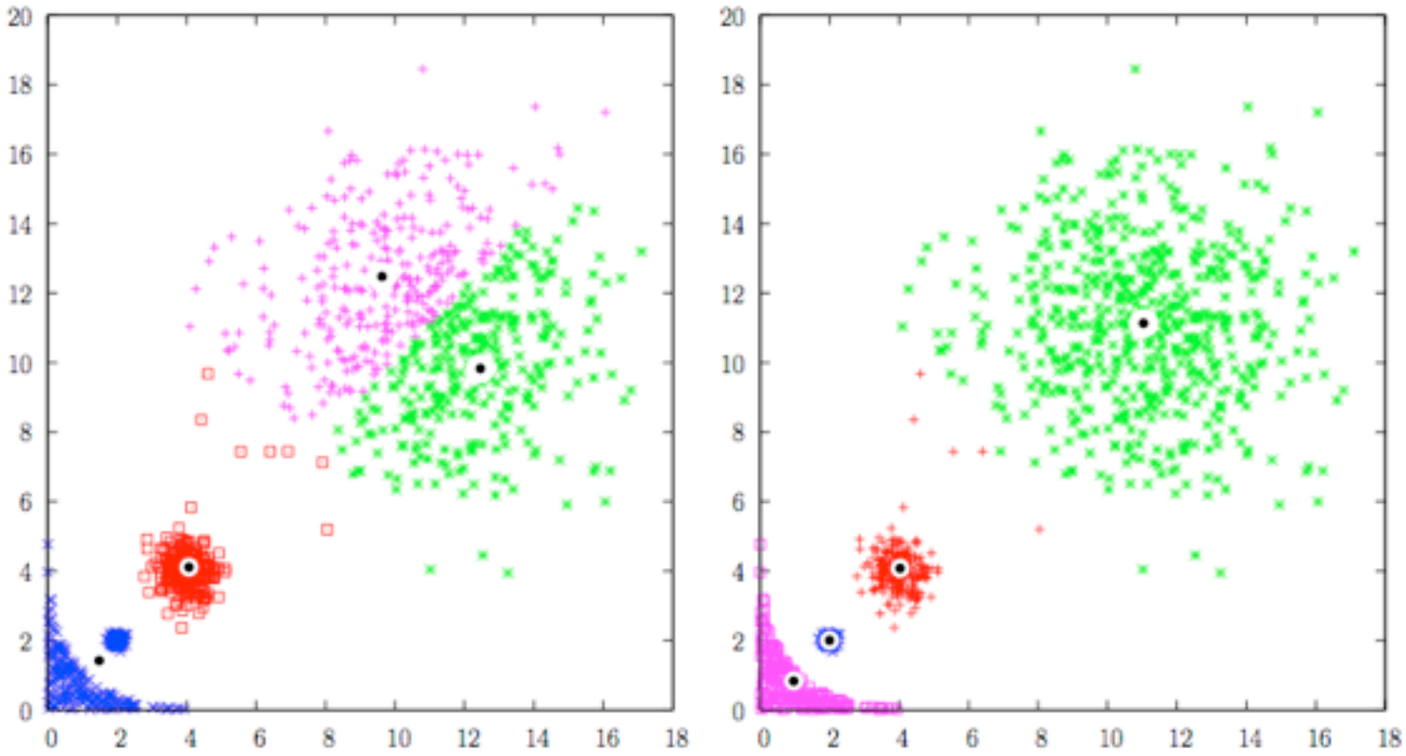


Fig. 1. Clusterings obtained by minimizing Euclidean (left) and Kullback-Leibler (right) potential. The centroids are shown as black dots.

Nock, Luosto, and Kivinen (2008). Mixed Bregman Clustering with Approximation Guarantees, p. 154–169 *in* Machine Learning and Knowledge Discovery in Databases (Daelemans, Goethals, and Morik, eds.), Springer.

Clustering partition subsets by Kullback-Leibler divergence

- Tricky—standard K-means does not work; need to use k-medians or k-medoids instead; had to write custom code.
- Preliminary results: No change for elasmobranch phylogeny (still get “wacky” tree)
- Method needs to be validated using computer simulation before I will be satisfied that it works (in progress).