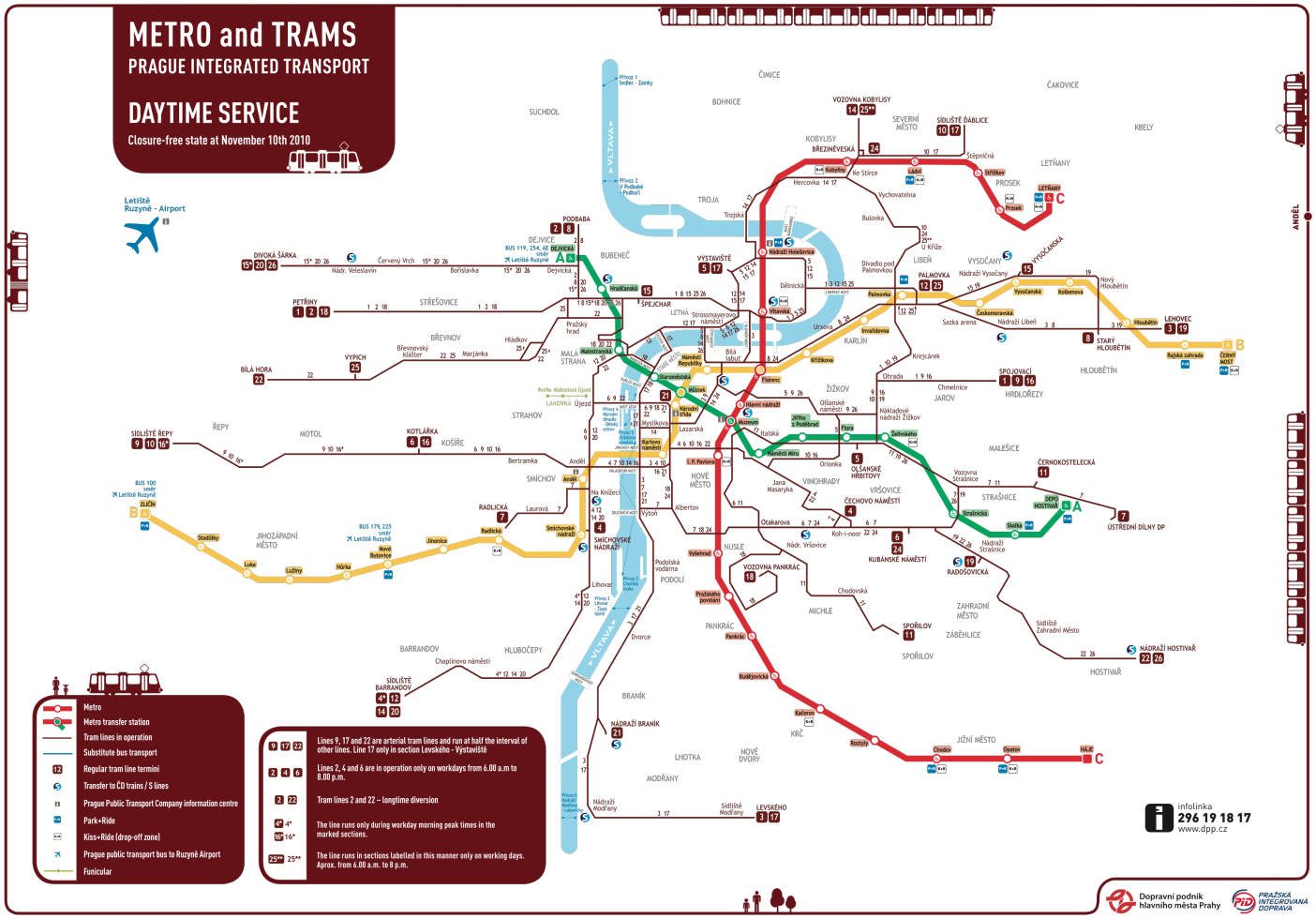
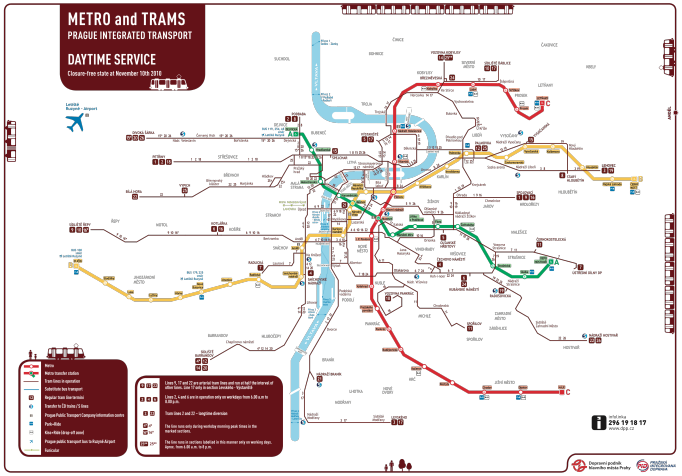
A model of the Prague Metro



Idea from Paul Lewis

A less complex model of the Prague Metro

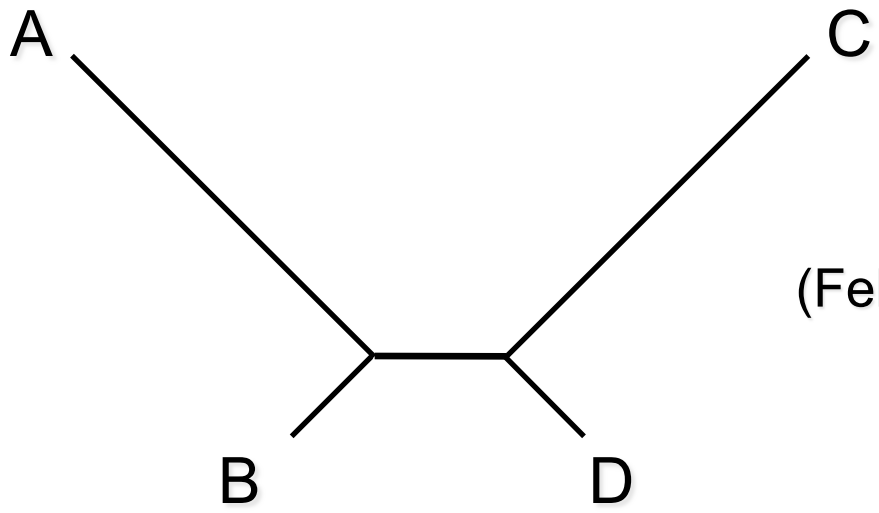
Model selection



Why do models matter?

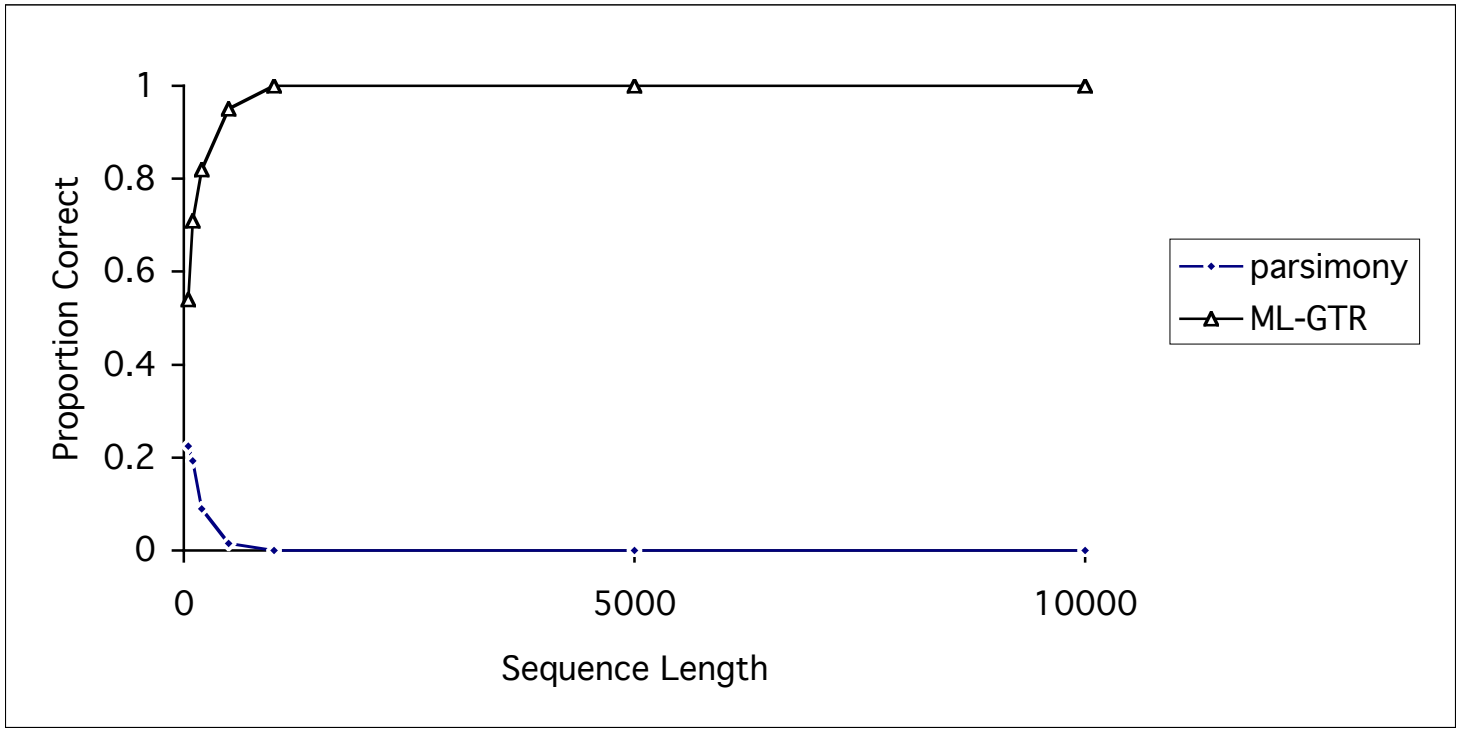
- Model-based methods including ML and Bayesian inference (typically) make a *consistent* estimate of the phylogeny (estimate converges to true tree as number of sites increases toward infinity)

  ... 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!

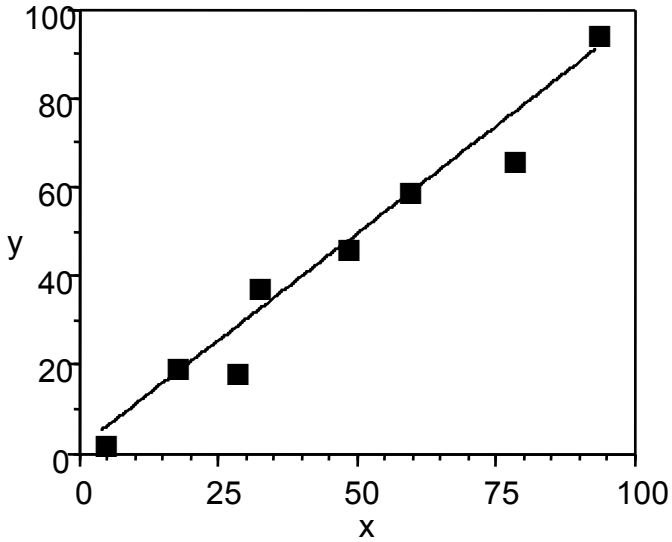
  Likelihood can also be inconsistent if the model is oversimplified

- 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?

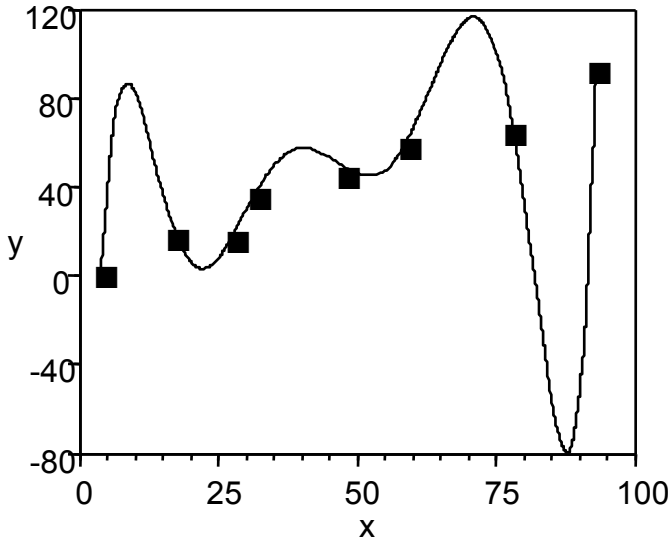
- A model that appropriately balances fit of the data with simplicity (parsimony, in a different sense)

  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**

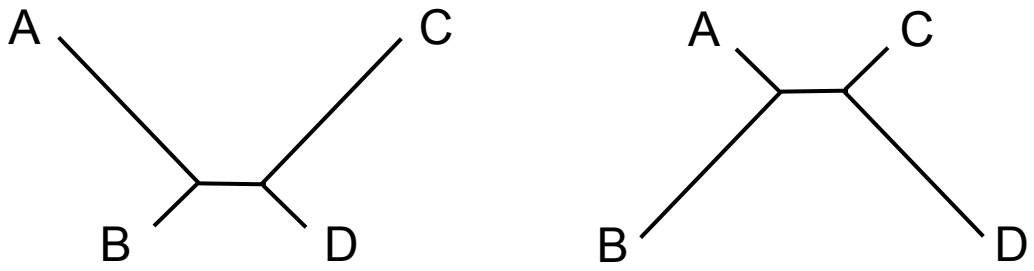
  - Parsimony lies between the evils of underfitting and overfitting. The concept of parsimony has a long history in in the sciences. Often this has been expressed as "Occam's razor"—shave away all that is not necessary. Parsimony in statistics represents a tradeoff between bias and variance as a function 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")

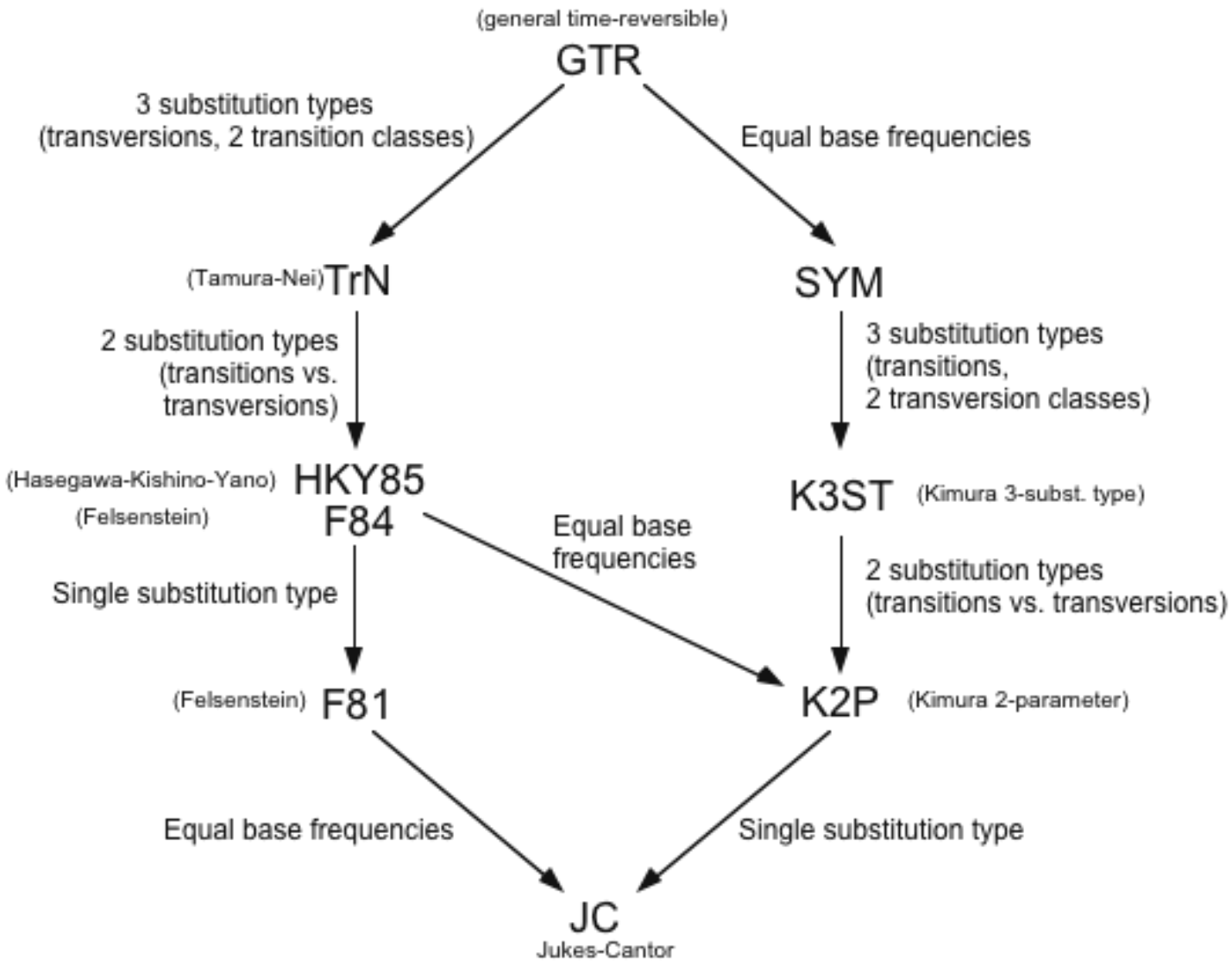
E.g.:



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



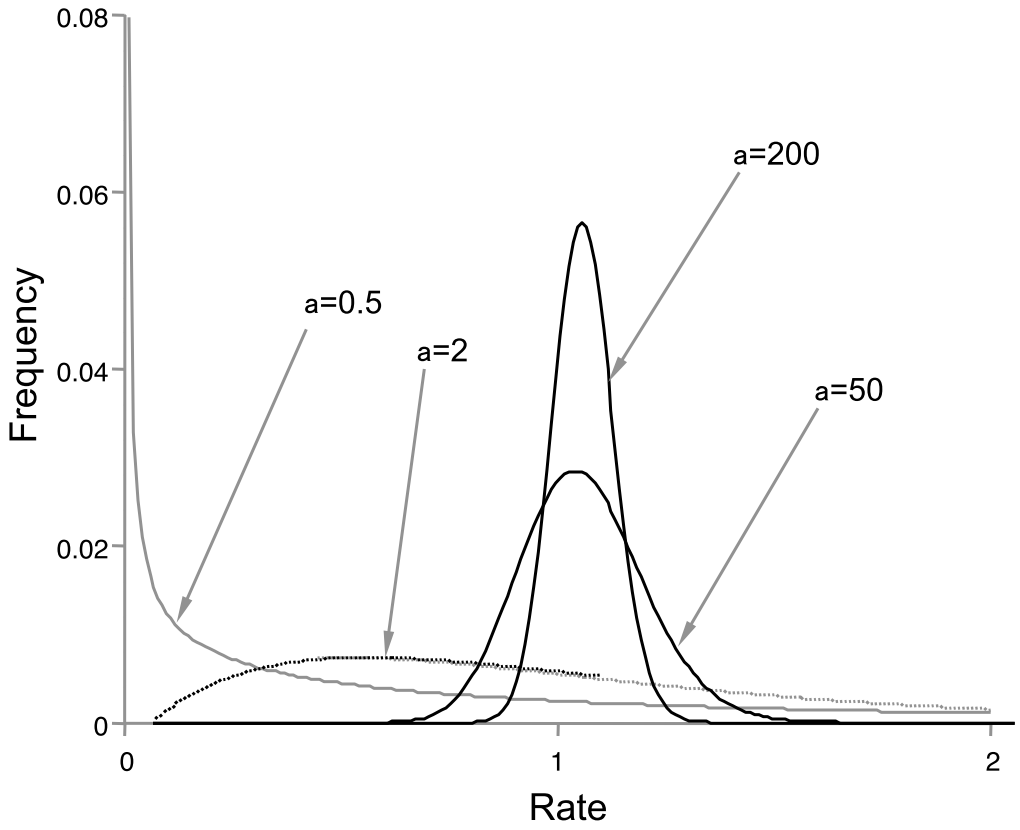
Among site rate heterogeneity

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
```

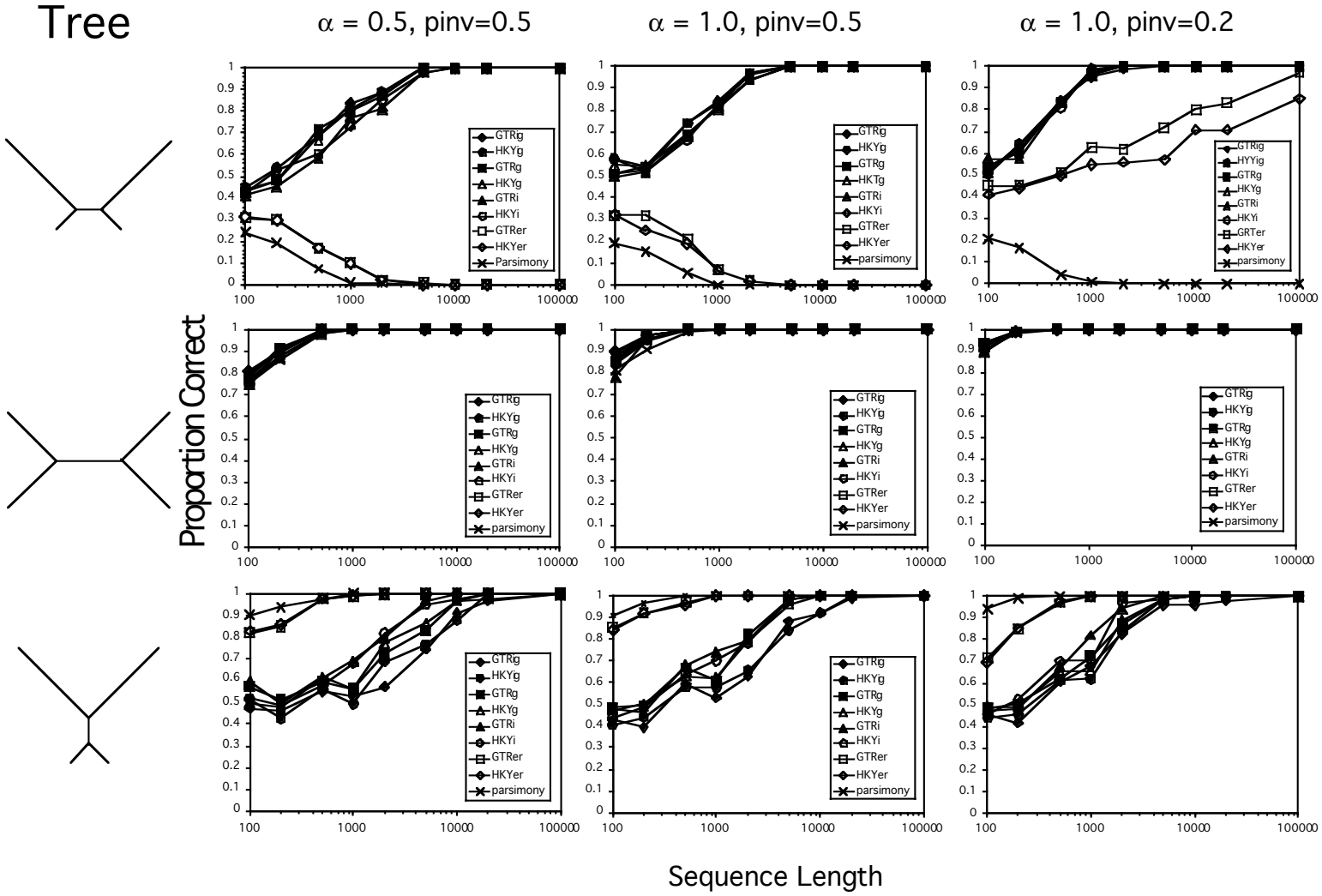
- Proportion of invariable sites
  - Some sites extremely unlikely to change due to strong functional or structural constraint (Hasegawa et al., 1985)
- Gamma-distributed rates
  - Rate variation assumed to follow a gamma distribution with shape parameter α
- Site-specific rates (another way to model ASRV)
  - Different relative rates assumed for pre-assigned subsets of sites

Modeling ASRV with gamma distribution



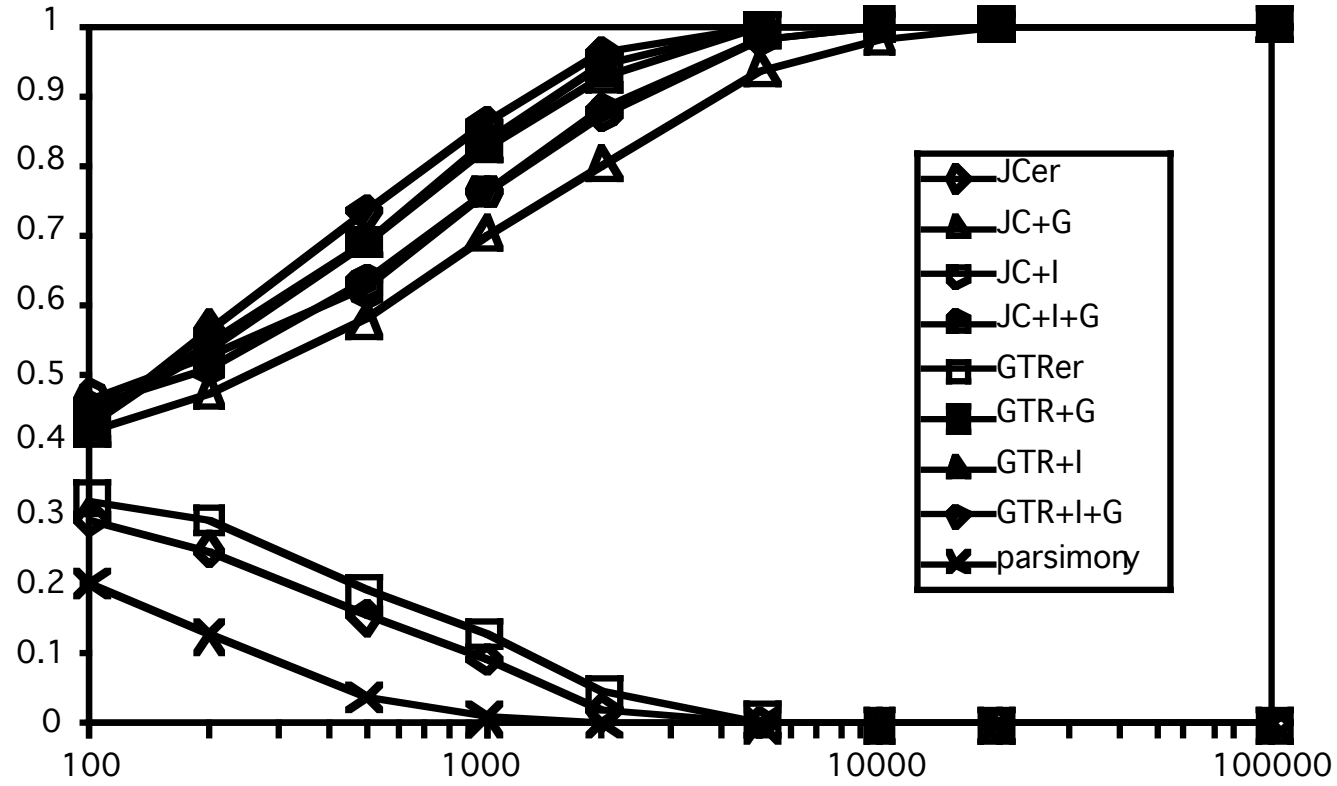
…can also include a proportion of “invariable” sites (p_{inv})

Performance of ML when its model is violated

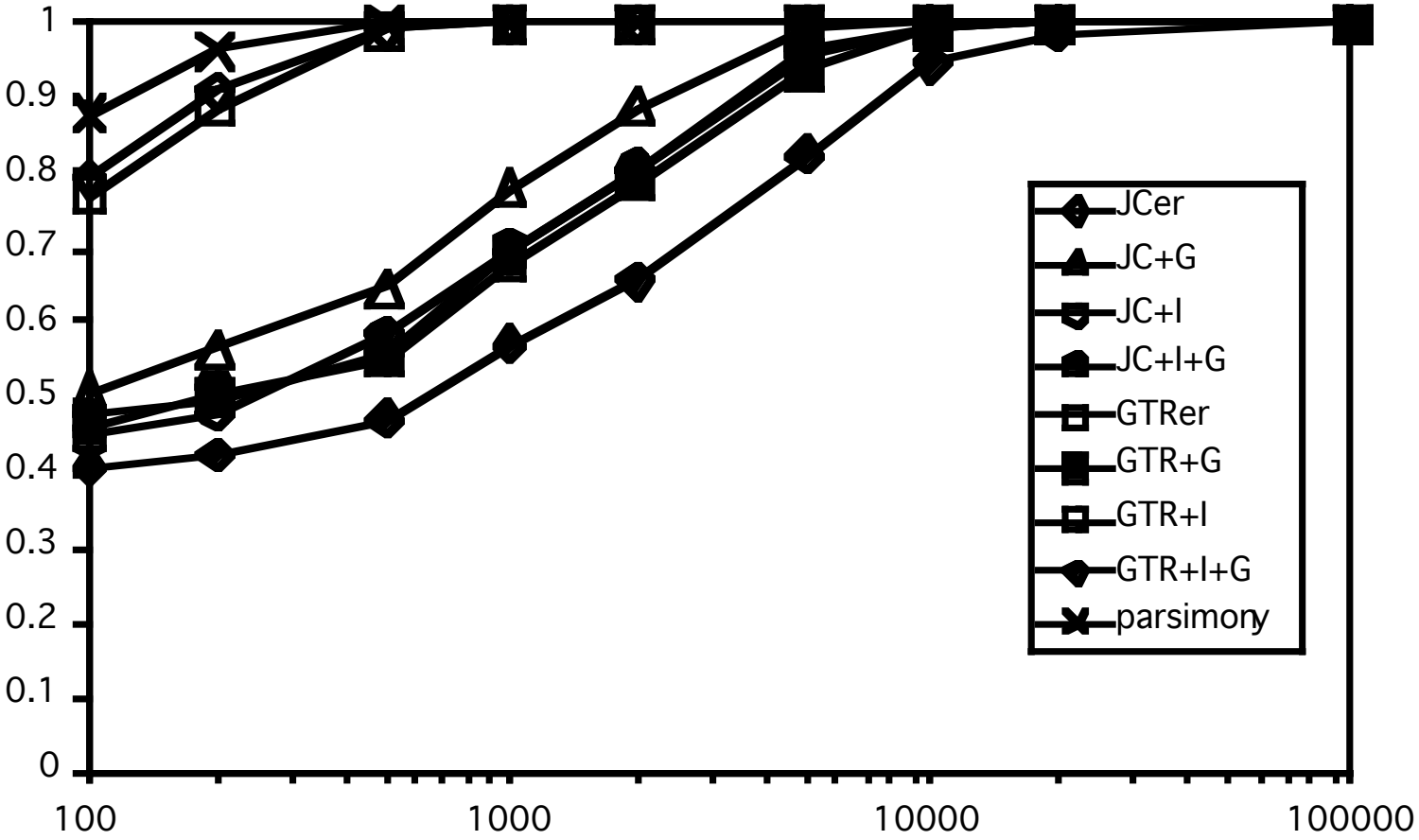


“MODERATE”–Felsenstein zone

$\alpha = 1.0, p_{inv}=0.5$



“MODERATE”–Inverse-Felsenstein zone



Model selection criteria

- Likelihood ratio tests

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

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

- Akaike information criterion (AIC)

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

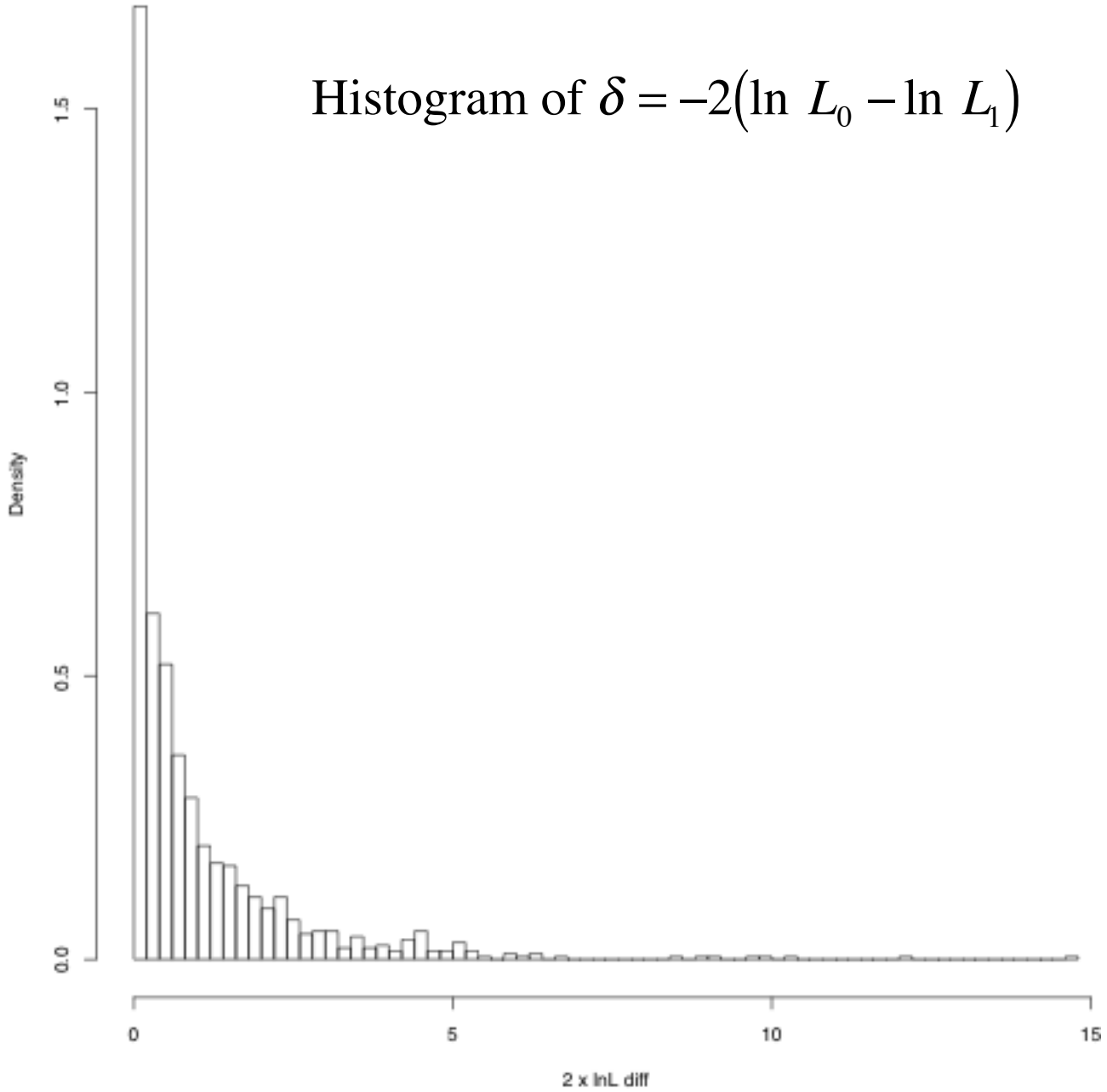
where *K* is the number of free parameters estimated

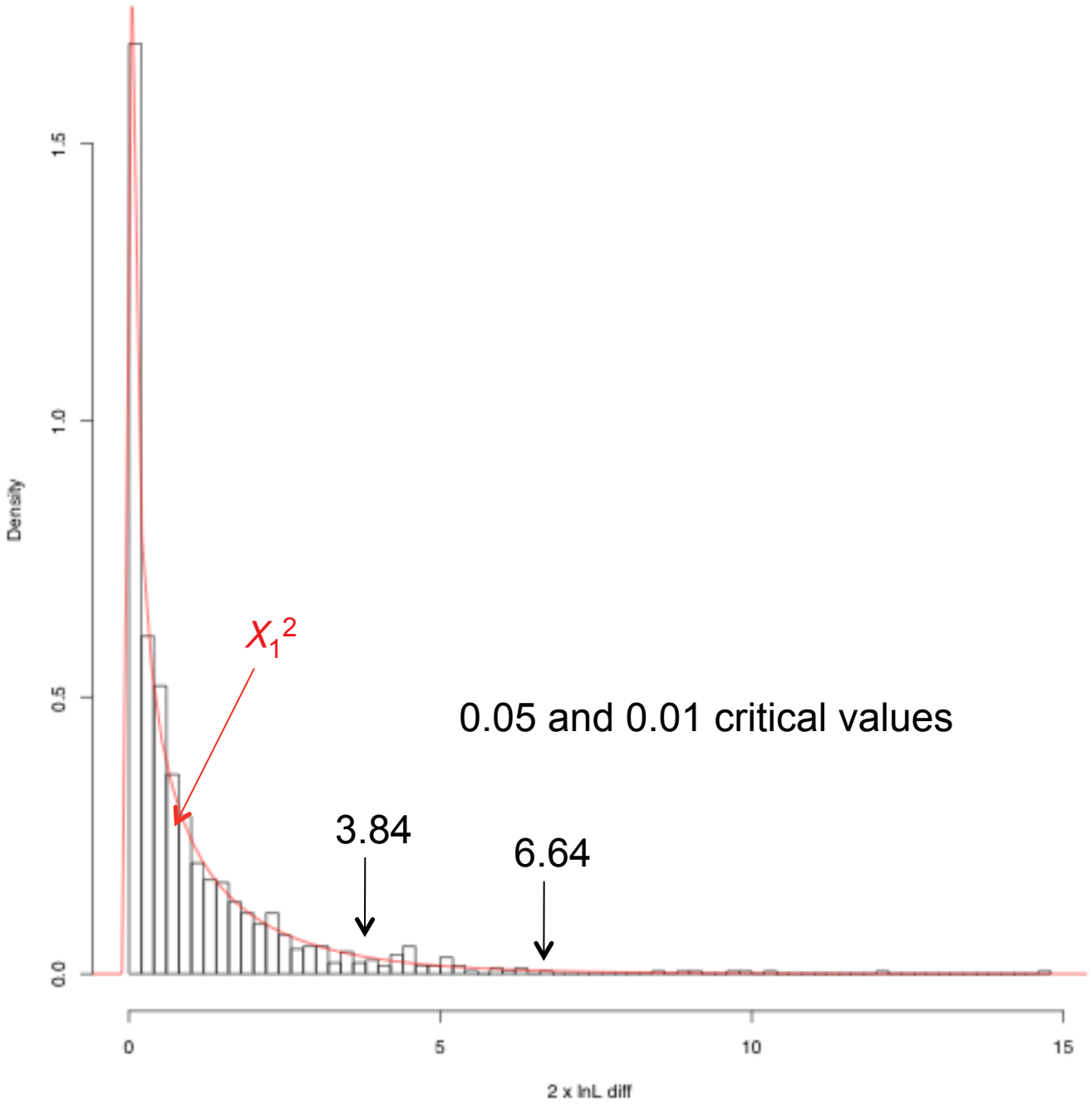
- Bayesian information criterion (BIC)

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

where *K* is the number of free parameters estimated and *n* is the "sample size" (typically number of sites)

Histogram of $\delta = -2\left(\ln\ L_0 - \ln\ L_1\right)$




Density
0.0
0.5
1.0
1.5
X_1^2
0.05 and 0.01 critical values
3.84
6.64
0
5
10
15
2 x lnL diff

Model selection and partitioning

- Partitioning schemes
  - By gene
  - By codon
  - By gene/codon combination
  - Stems vs. loops
  - Coding vs. noncoding
  - Other clustering methods
- Overpartitioning is a risk

  Slightly silly example (different variations on the JC model):
  - Gene A: HKY+G, π = (0.26,0.24,0.23,0.27), kappa=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 (p_{inv}=0.05)
- Use PartitionFinder (http://www.robertlanfear.com/partitionfinder/)