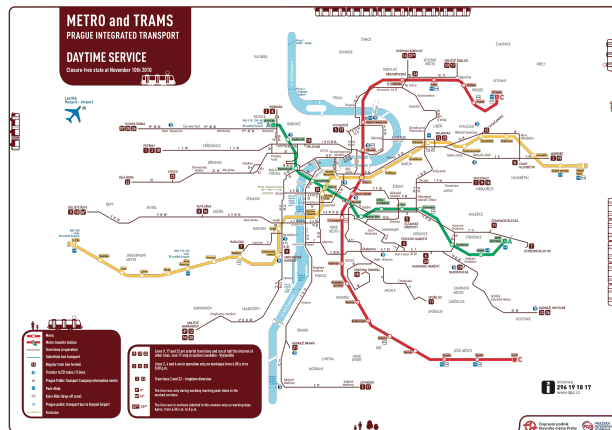# Which is more useful?

"Reality"

Detailed map

Detailed public transportation

Simplified metro

Models don't need to reflect reality

- A **model** is an intentional simplification of a complex situation designed to eliminate extraneous detail in order to focus attention on the essentials of the situation. (Daniel L. Hartl, 2000)

- "The most that can be expected from any model is that it can supply a useful approximation to reality: **All models are wrong; some models are useful"**. (George E. P. Box, 1987)

- A model is a simplification or approximation of reality and hence will not reflect all of reality ... While a model can never be "truth," a model might be ranked from very useful, to useful, to somewhat useful to, finally, essentially useless. (Burnham and Anderson, 2002)

- Model selection is a process of seeking the least inadequate model from a predefined set, all of which may be grossly inadequate as a representation of reality. (J. J. Welch, 2006)
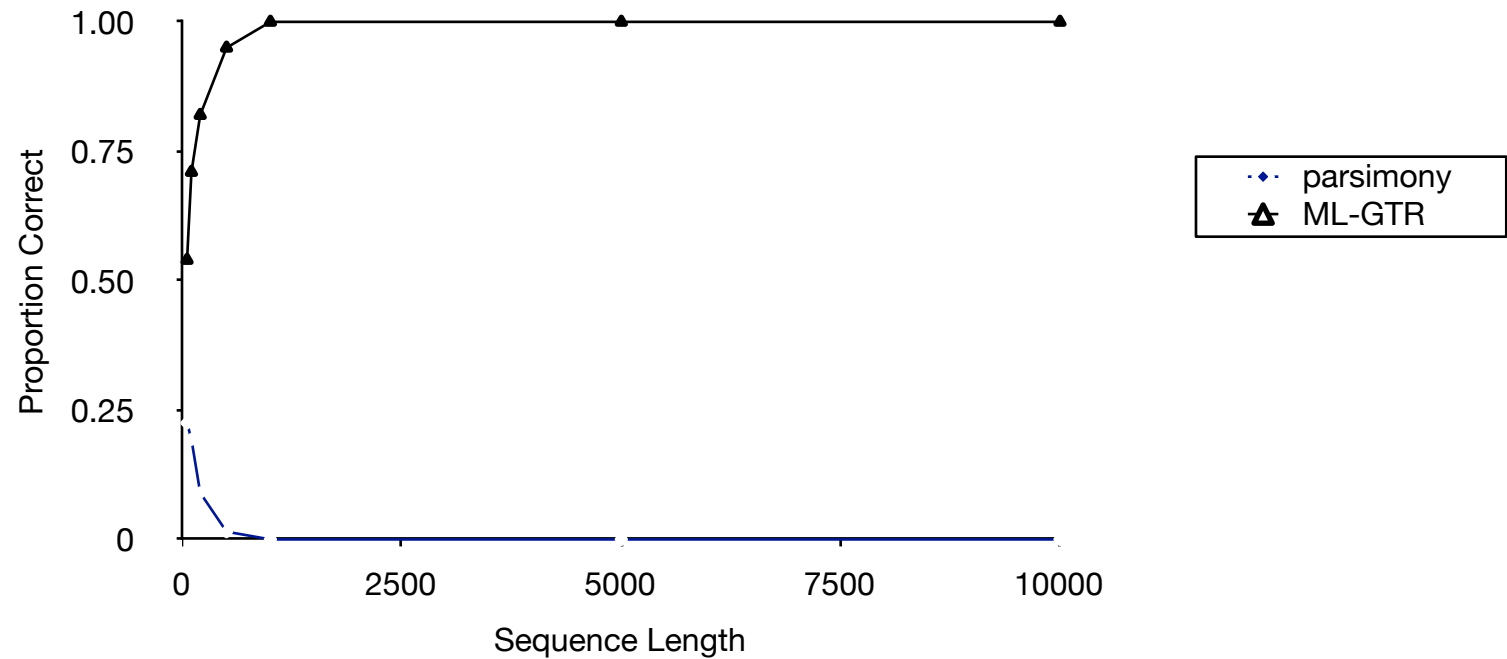
Why do models matter?

- 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"

A   C

(Felsenstein, 1978)

B   D
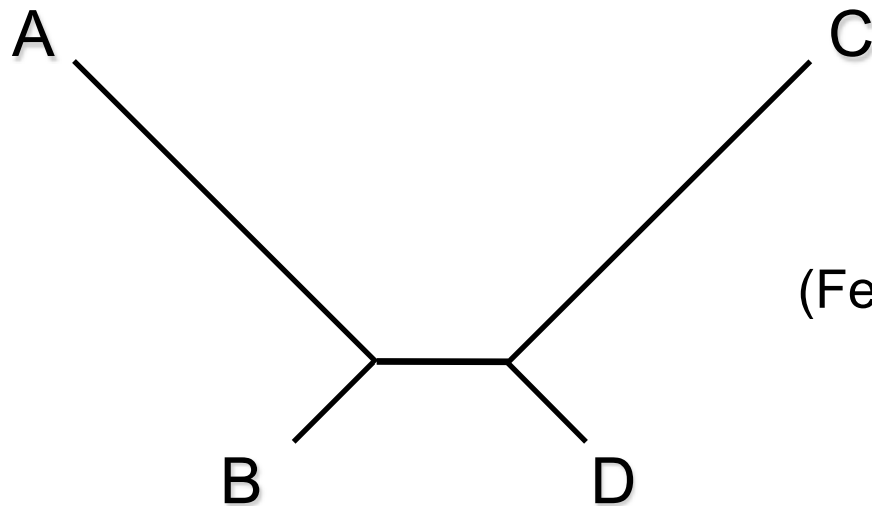
In the Felsenstein Zone

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
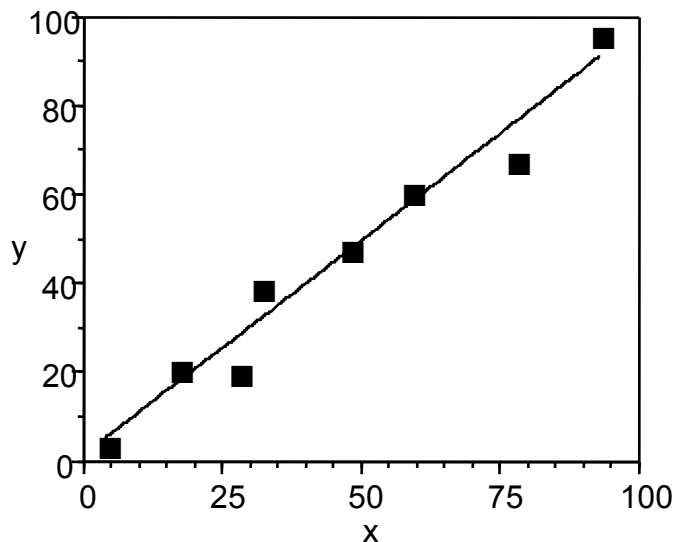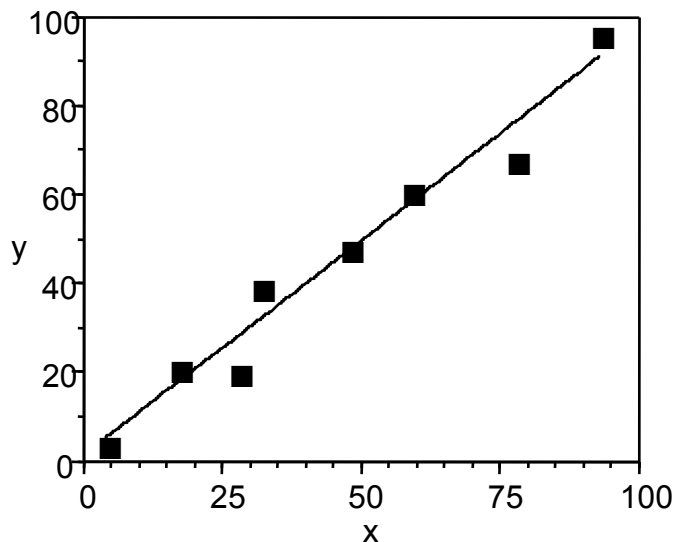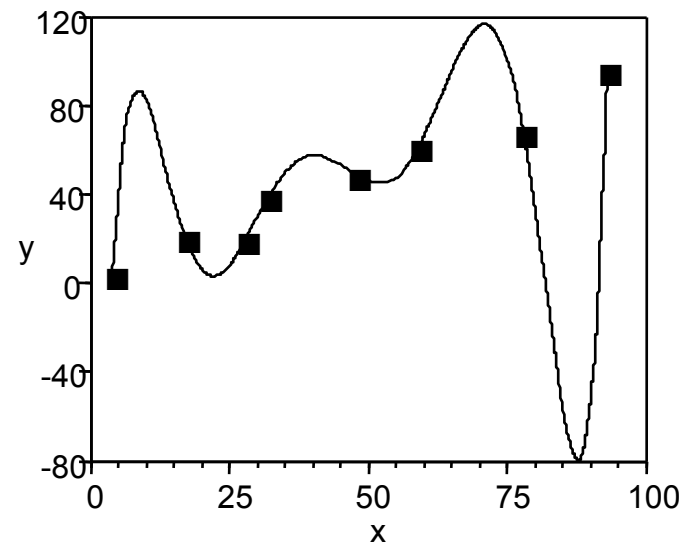

$$y = 1.30 + 0.965x$$
$$(r^2 = 0.963)$$
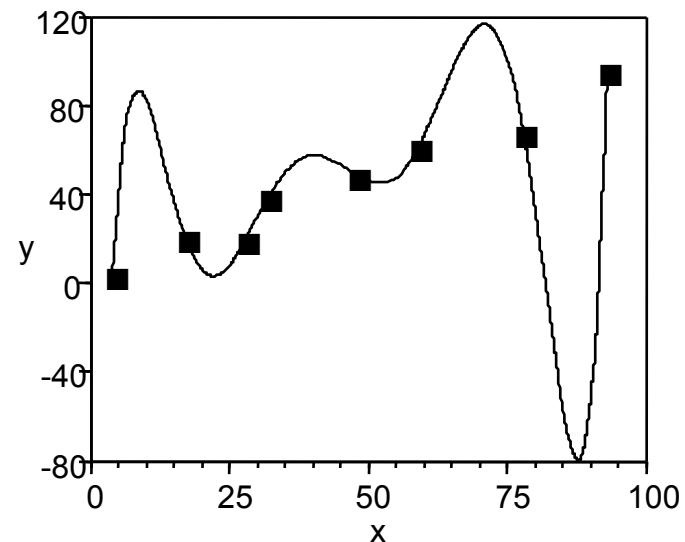
$$y = -330 + 134x - 15.5x^2 + 0.816x^3$$
$$- 0.0225x^4 + 0.000335x^5$$
$$- 0.00000255x^6 + 0.00000000777x^7$$
$$(r^2 = 1.000)$$

What is a "good" model?

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. (Burnham and Anderson, 1998)


$$y = 1.30 + 0.965x$$
$$(r^2 = 0.963)$$

$$y = -330 + 134x - 15.5x^2 + 0.816x^3$$
$$- 0.0225x^4 + 0.000335x^5$$
$$- 0.00000255x^6 + 0.00000000777x^7$$
$$(r^2 = 1.000)$$
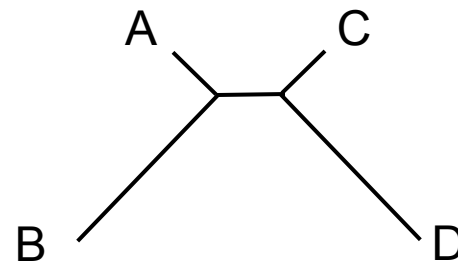
Why models don't have to be perfect

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")
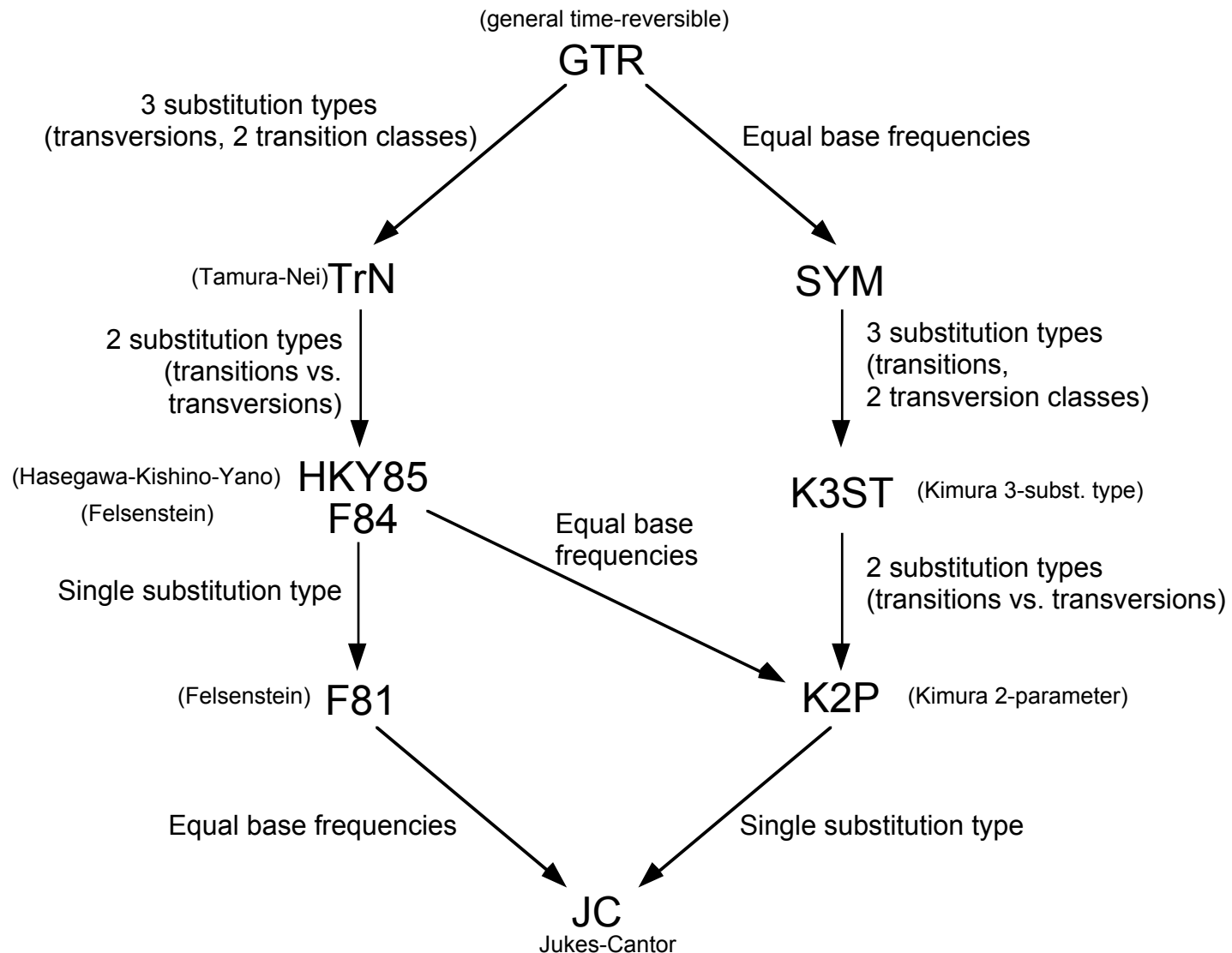
E.g.:
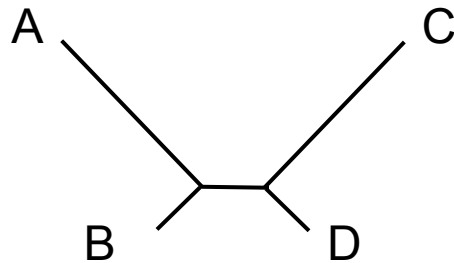
Half of sites   Other half

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) ("heterotachy")

GTR Family of Reversible DNA Substitution Models

(general time-reversible)

GTR

3 substitution types
(transversions, 2 transition classes)

Equal base frequencies

(Tamura-Nei) **TrN**

SYM

2 substitution types
(transitions vs.
transversions)

3 substitution types
(transitions,
2 transversion classes)

(Hasegawa-Kishino-Yano) **HKY85**

(Felsenstein) **F84**

K3ST (Kimura 3-subst. type)

Equal base
frequencies

Single substitution type

2 substitution types
(transitions vs. transversions)

(Felsenstein) **F81**

K2P (Kimura 2-parameter)

Equal base frequencies

Single substitution type

JC
Jukes-Cantor

Monday, January 30, 17

Among site rate heterogeneity

```
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
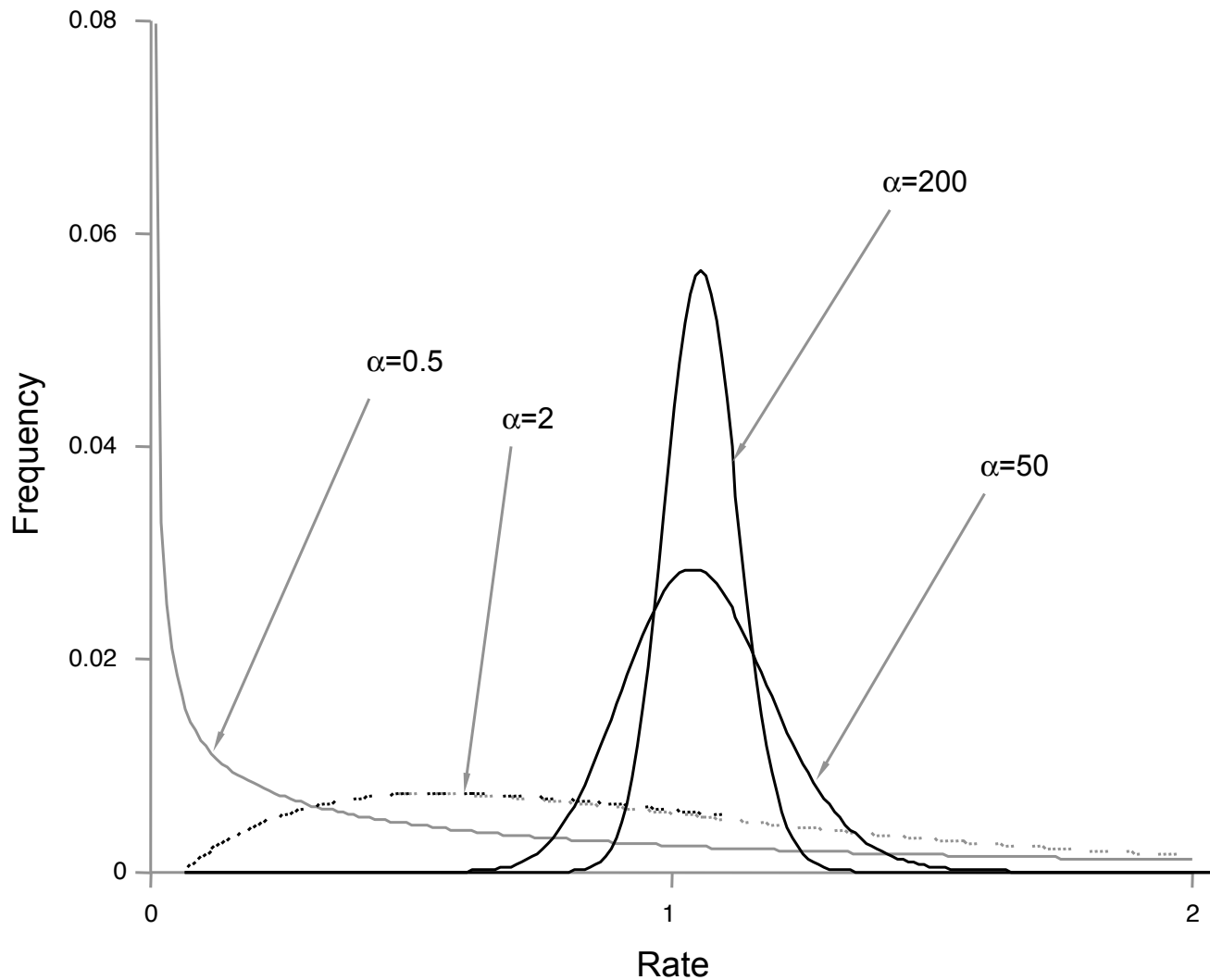
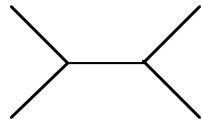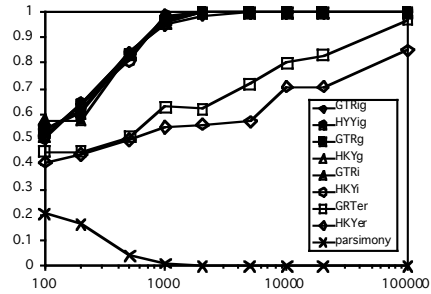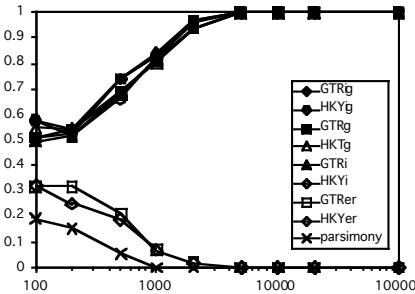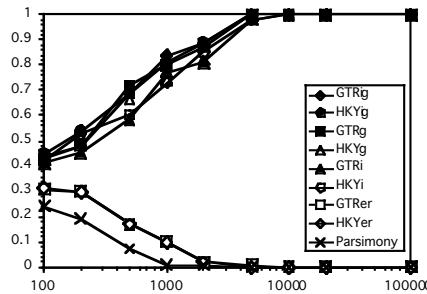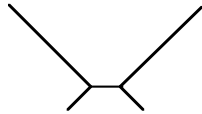...can also include a proportion of "invariable" sites (p_{inv})

Performance of ML when its model is violated

Tree

α = 0.5, pinv=0.5   α = 1.0, pinv=0.5   α = 1.0, pinv=0.2

Proportion Correct
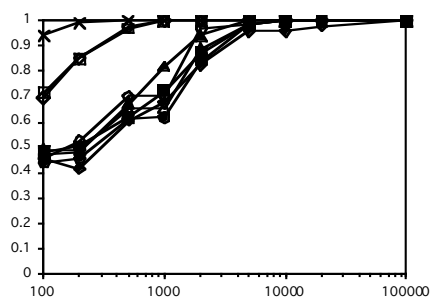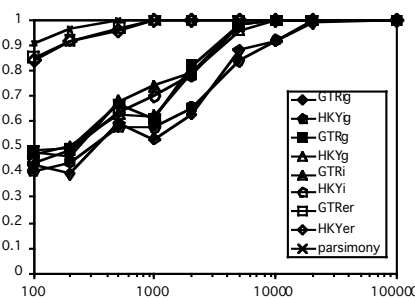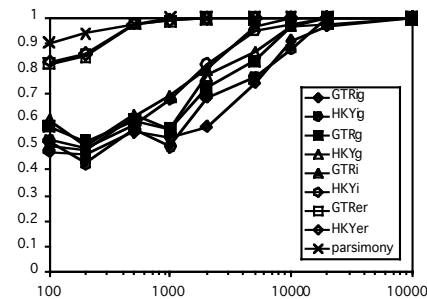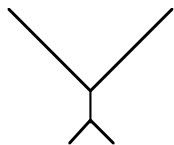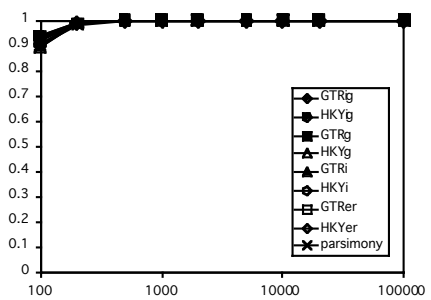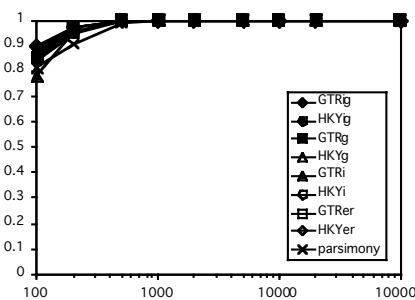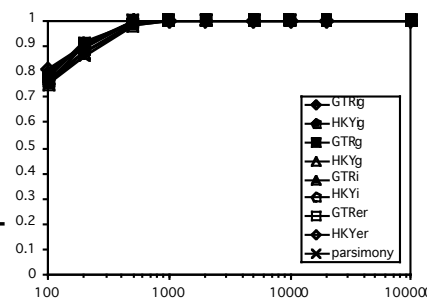
Sequence Length

"MODERATE"–Felsenstein zone

$$\alpha = 1.0,\ p_{inv}=0.5$$
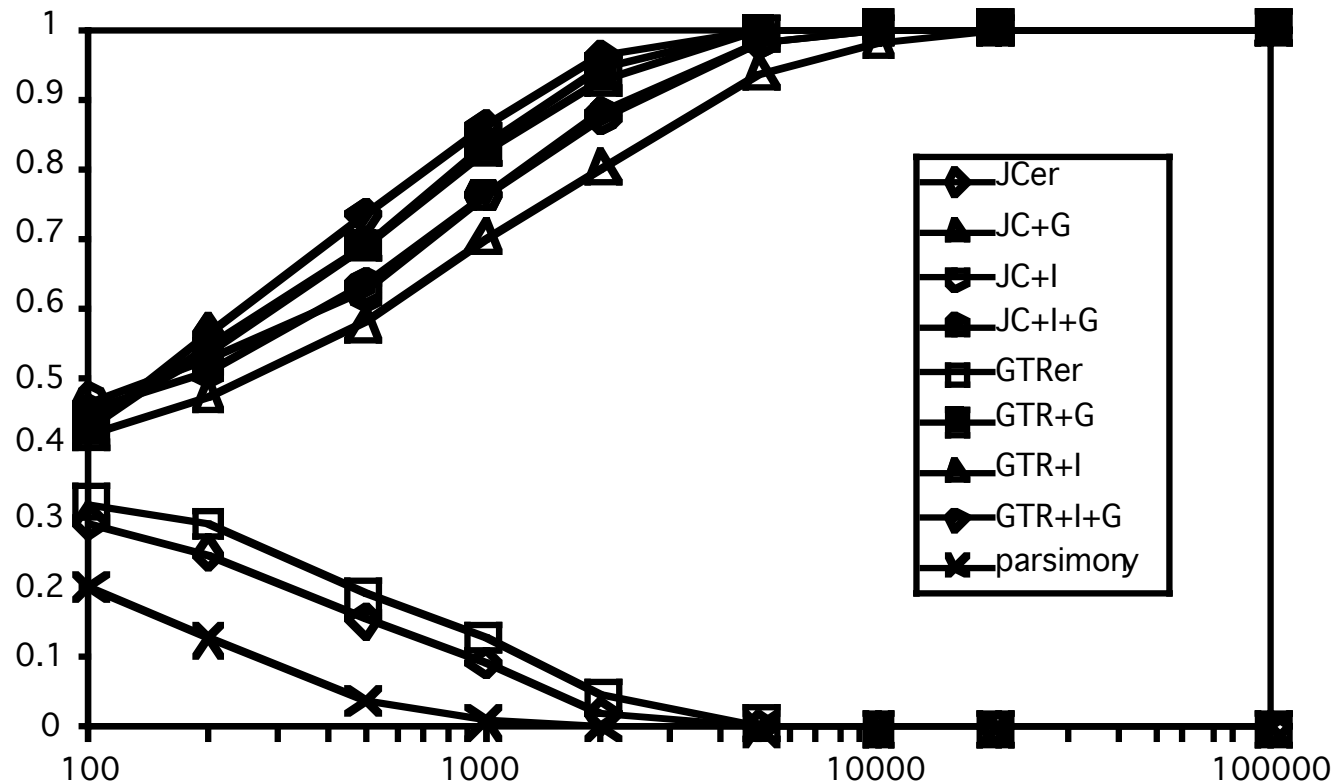
"MODERATE"–Inverse-Felsenstein zone
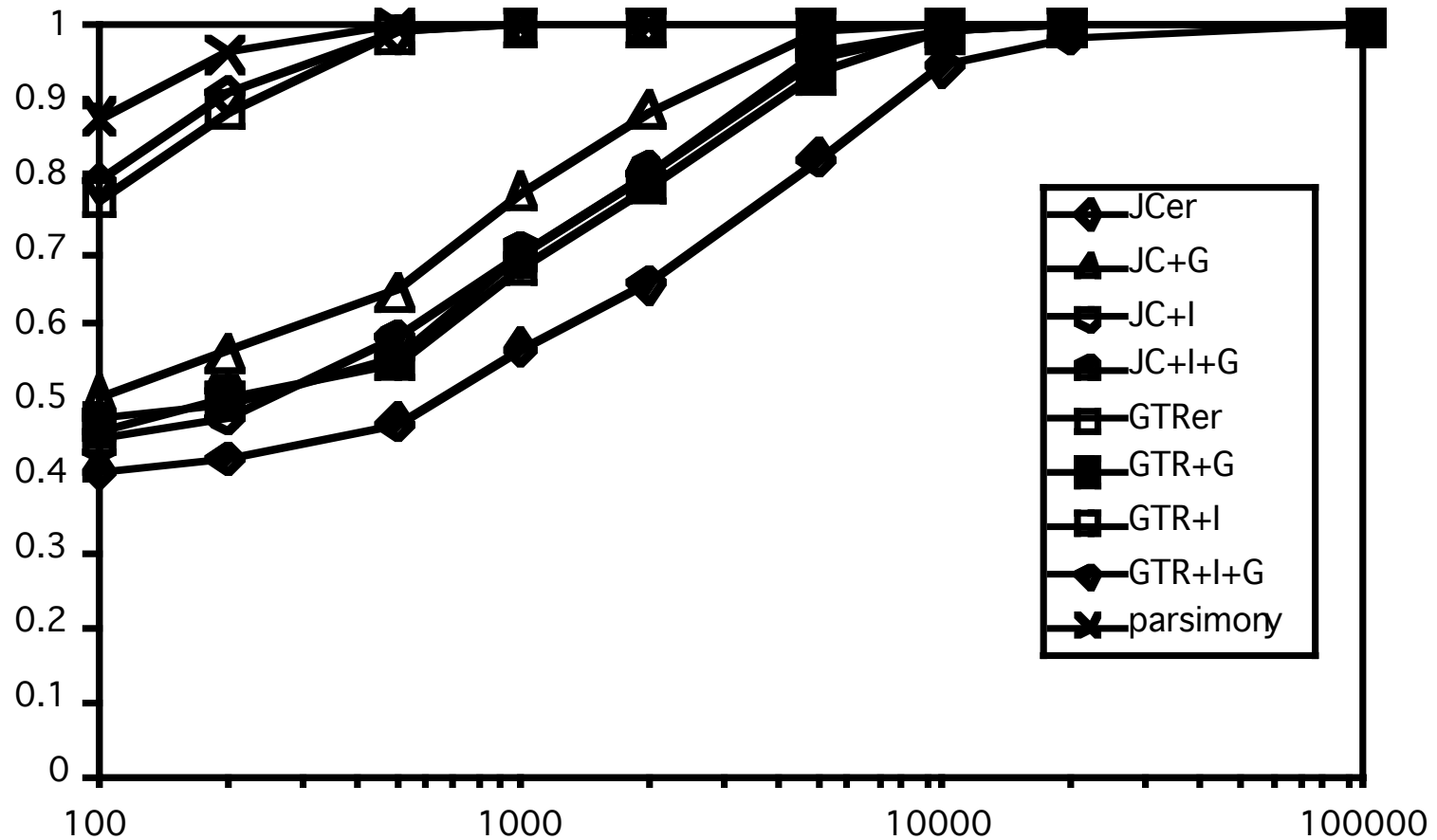
Model selection criteria

- ## 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

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

JC vs K80 models

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

JC vs K80 models

X^2

0.05 and 0.01 critical values

3.84

6.64
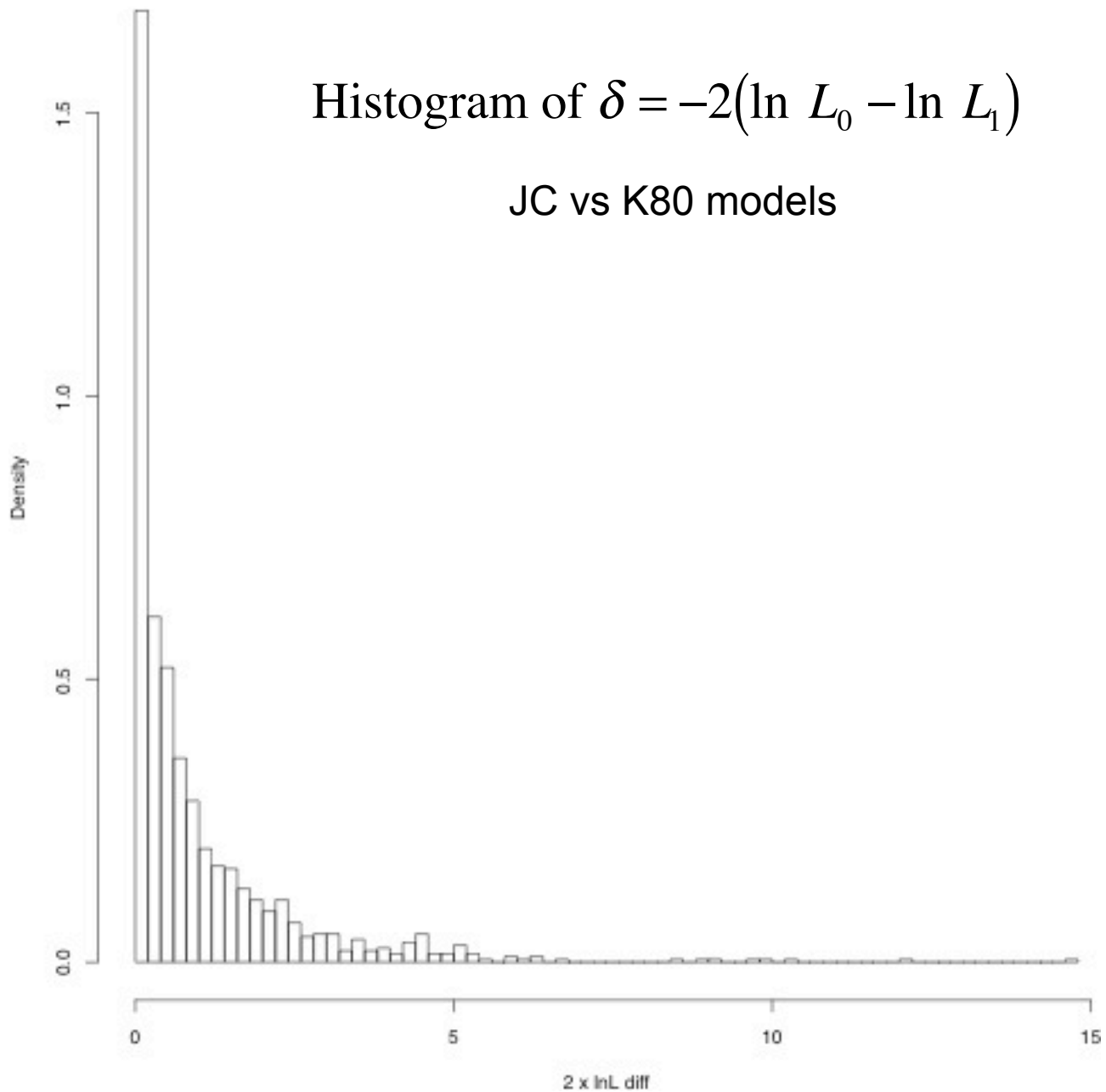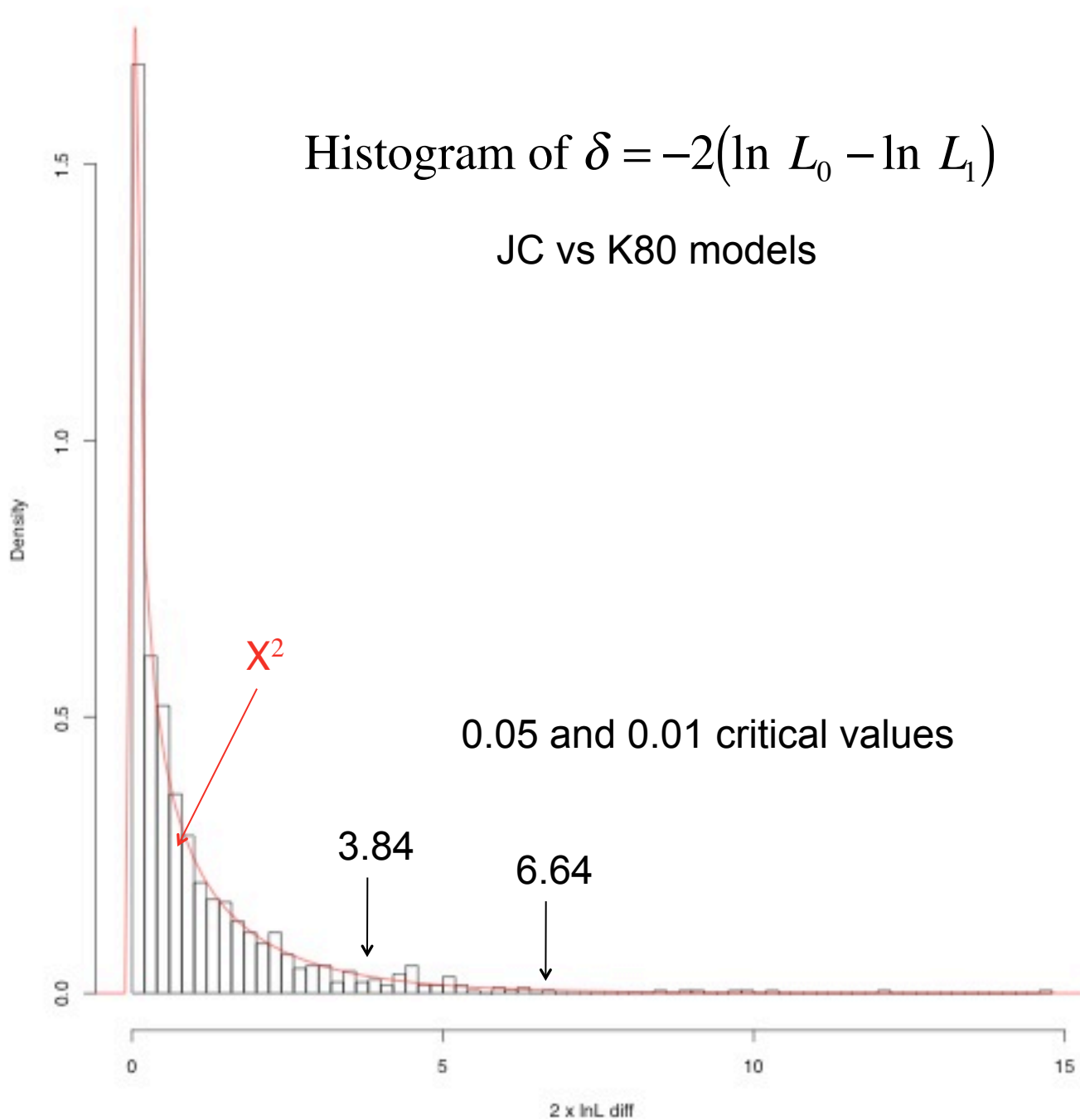
Model selection criteria

- ## Akaike information criterion (AIC)

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

where K is the number of free parameters estimated

- ## AICc (corrected AIC)

- ## 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)

AIC vs. BIC

– BIC performs well when true model is contained in model set, and among a set of simple models, AIC often selects a more complex model than the truth (indeed, AIC is formally statistically inconsistent)

– But in phylogenetics, no model is as complex as the truth, and the true model will never be contained in the model set.

– BIC often chooses models that seem *too* simple, however.

jModelTest example

- java -jar jmodeltest-2.1.10/jModelTest.jar -d primate-mtDNA.nex -s 7 -t fixed -g 4 -f -i -AIC -BIC

20

Partitioned Models

Many authors have emphasized the importance of modeling heterogeneity among genes or other subsets of the data appropriately

> "...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:

Ⓐ Ⓑ Ⓒ

(A B) Ⓒ

(A C) Ⓑ

(B C) Ⓐ

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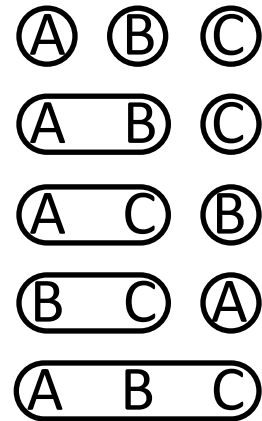
For each partitioning scheme, evaluate some set of models from the GTR+I+G (e.g., 56 models) according to AIC or BIC

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}

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

Greedy algorithm when there are too many
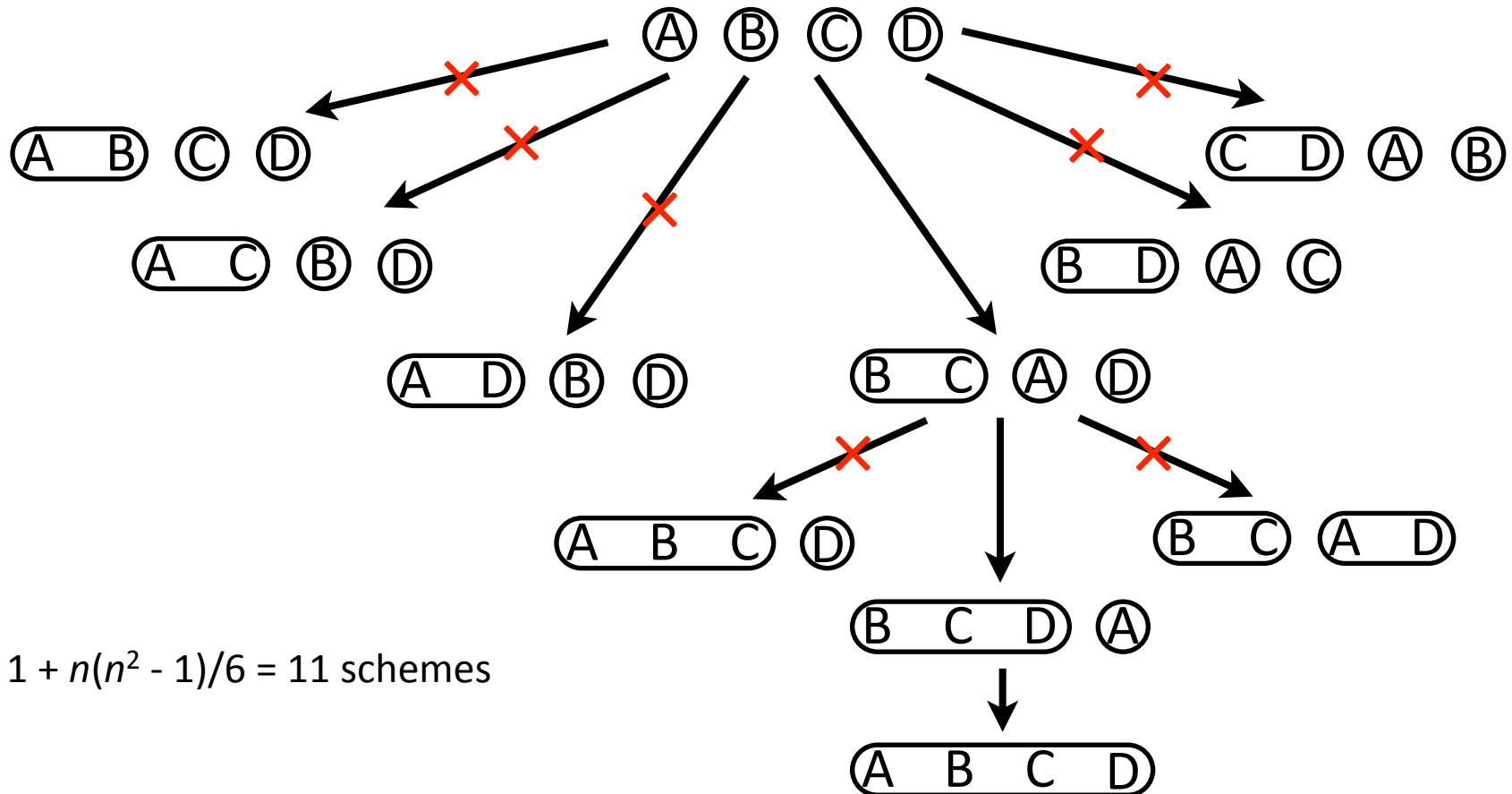
$1 + n(n^2 - 1)/6 = 11$ schemes

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

Monday, January 30, 17

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

Cluster analysis

- Li, Lu, and Orti (2008)

  Estimate model parameters on 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. Must use same model for all subsets.

- Lanfear et al. (PartitionFinder2)

  Hierarchical (or non-hierarchical kmeans) clustering using same idea as Li et al. (very efficient implementation)