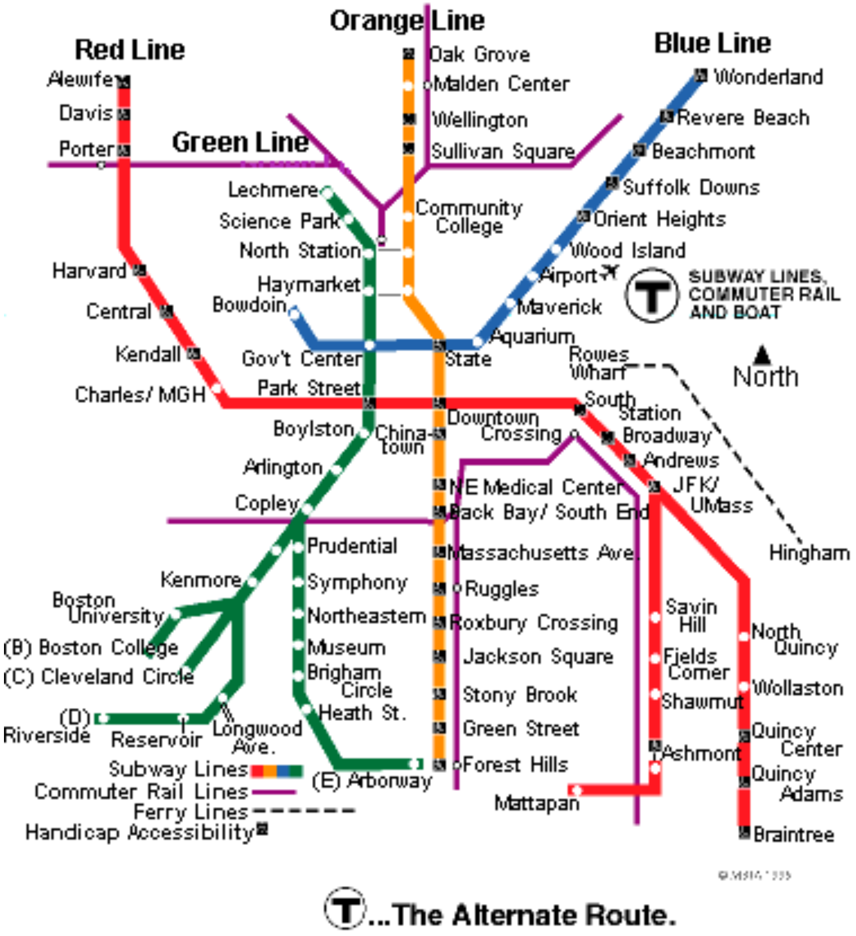
A very *practical* MBTA subway map



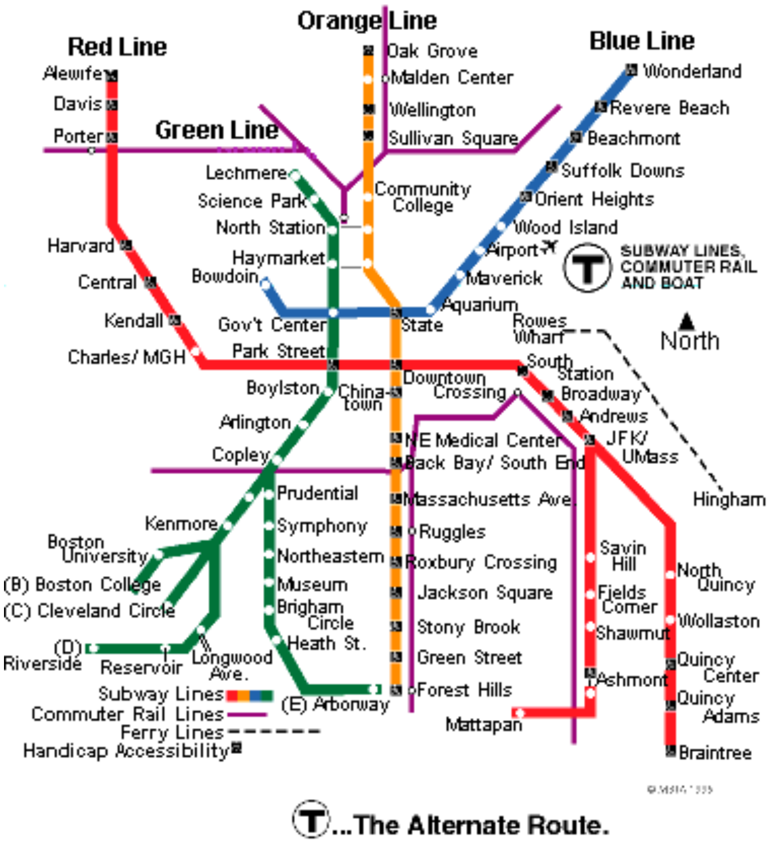
MBTA

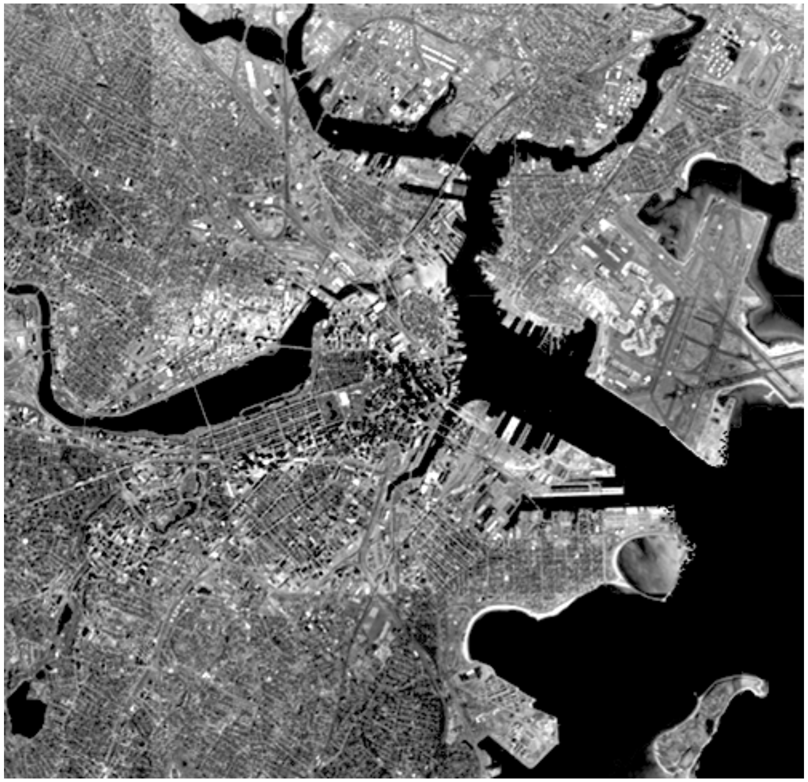
MASSACHUSETTS BAY TRANSPORTATION AUTHORITY

A very *realistic* MBTA subway map

MBTA
MASSACHUSETTS BAY TRANSPORTATION AUTHORITY

Which is more useful when you are trying to figure out how many stops there are between you and your destination?

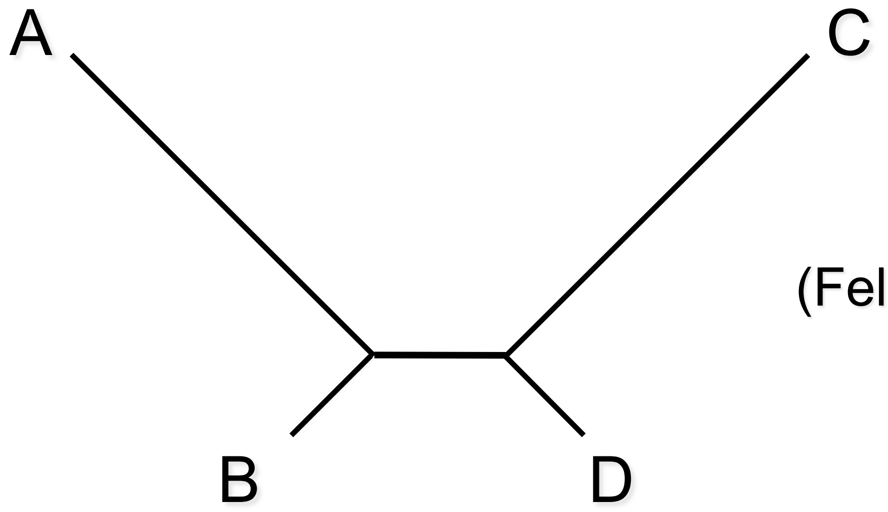




Why do models matter?

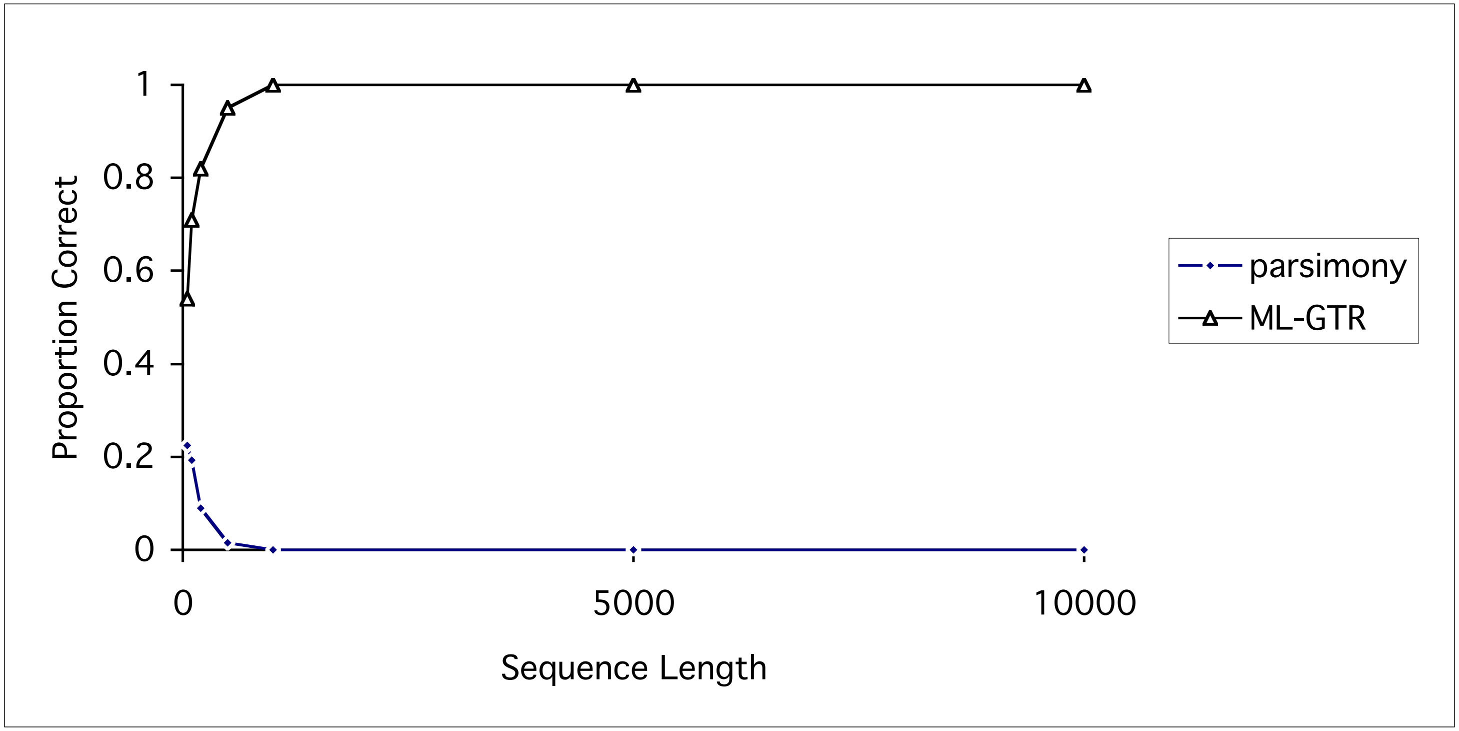
- Model-based methods including ML and Bayesian inference (typically) make a *consistent* estimate of the phylogeny (estimate converges to true tree as number of sites increases toward infinity)

  ... 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!

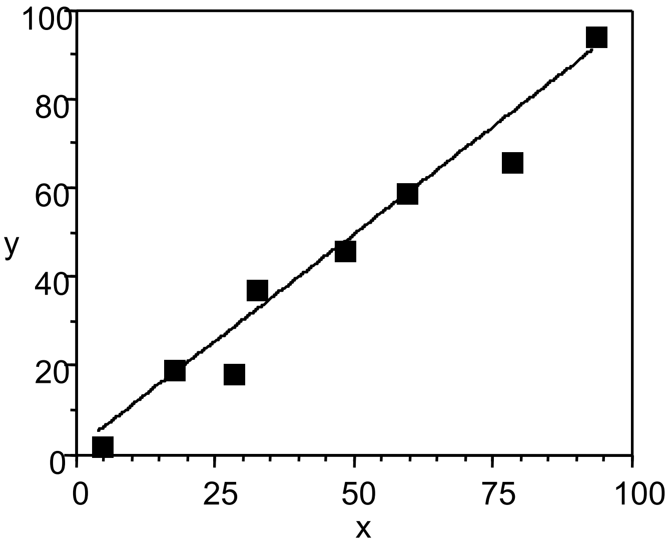
Likelihood can also be inconsistent if the model is oversimplified

- 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?

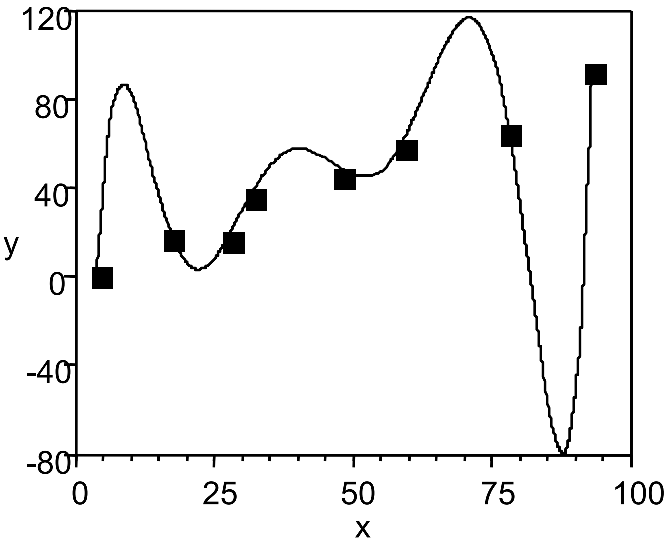
- A model that appropriately balances fit of the data with simplicity (parsimony, in a different sense)

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



$$y = 1.30 + 0.965x$$

$$(r^2 = 0.963)$$



$$\begin{aligned} y = &- 330 + 134x - 15.5x^2 + 0.816x^3 \\ &- 0.0225x^4 + 0.000335x^5 \\ &- 0.00000255x^6 + 0.00000000777x^7 \end{aligned}$$

$$(r^2 = 1.000)$$

“The Principle of Parsimony” in the world of statistics

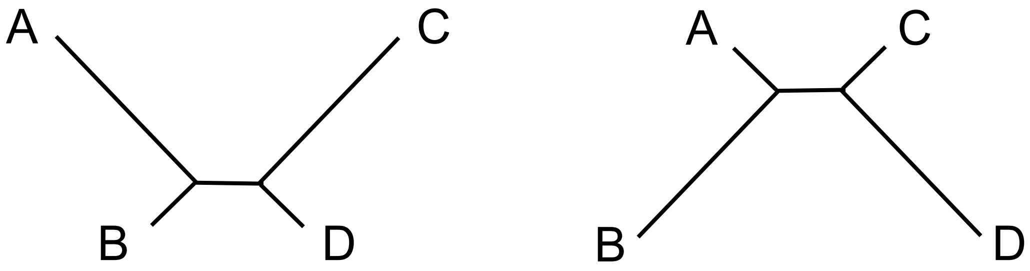
- **Burnham and Anderson (1998): Model Selection and Inference**
  - Parsimony lies between the evils of underfitting and overfitting. The concept of parsimony has a long history in in the sciences. Often this has been expressed as “Occam’s razor”—shave away all that is not necessary. Parsimony in statistics 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.:

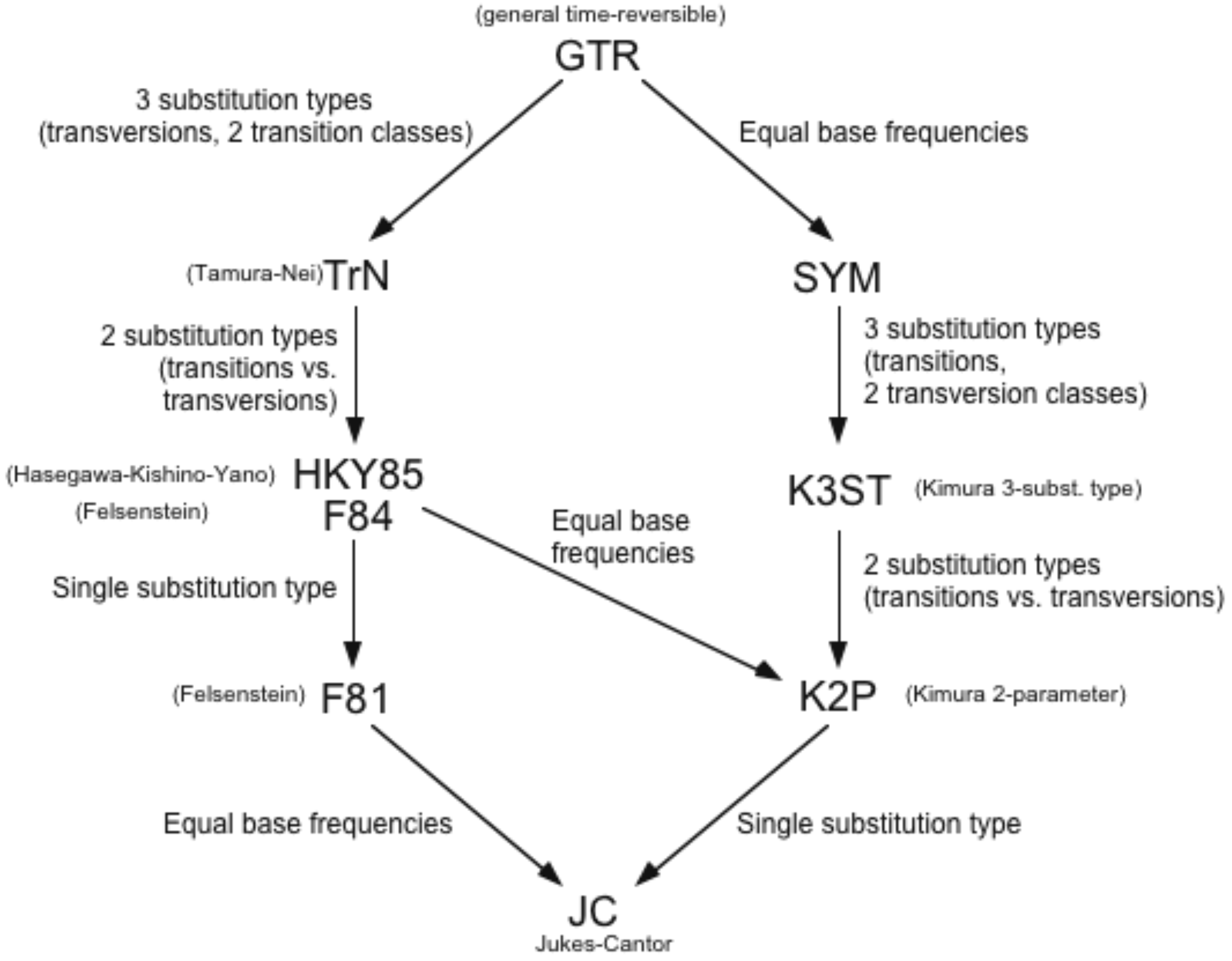


Half of sites

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



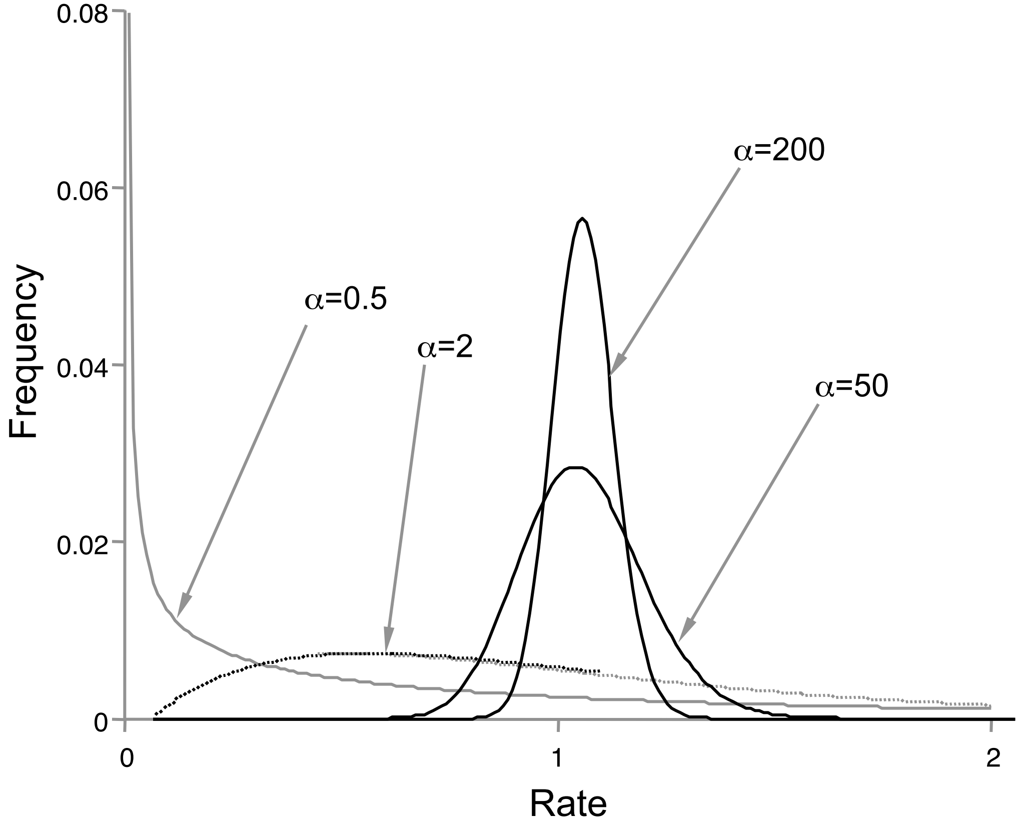
Among site rate heterogeneity

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
```

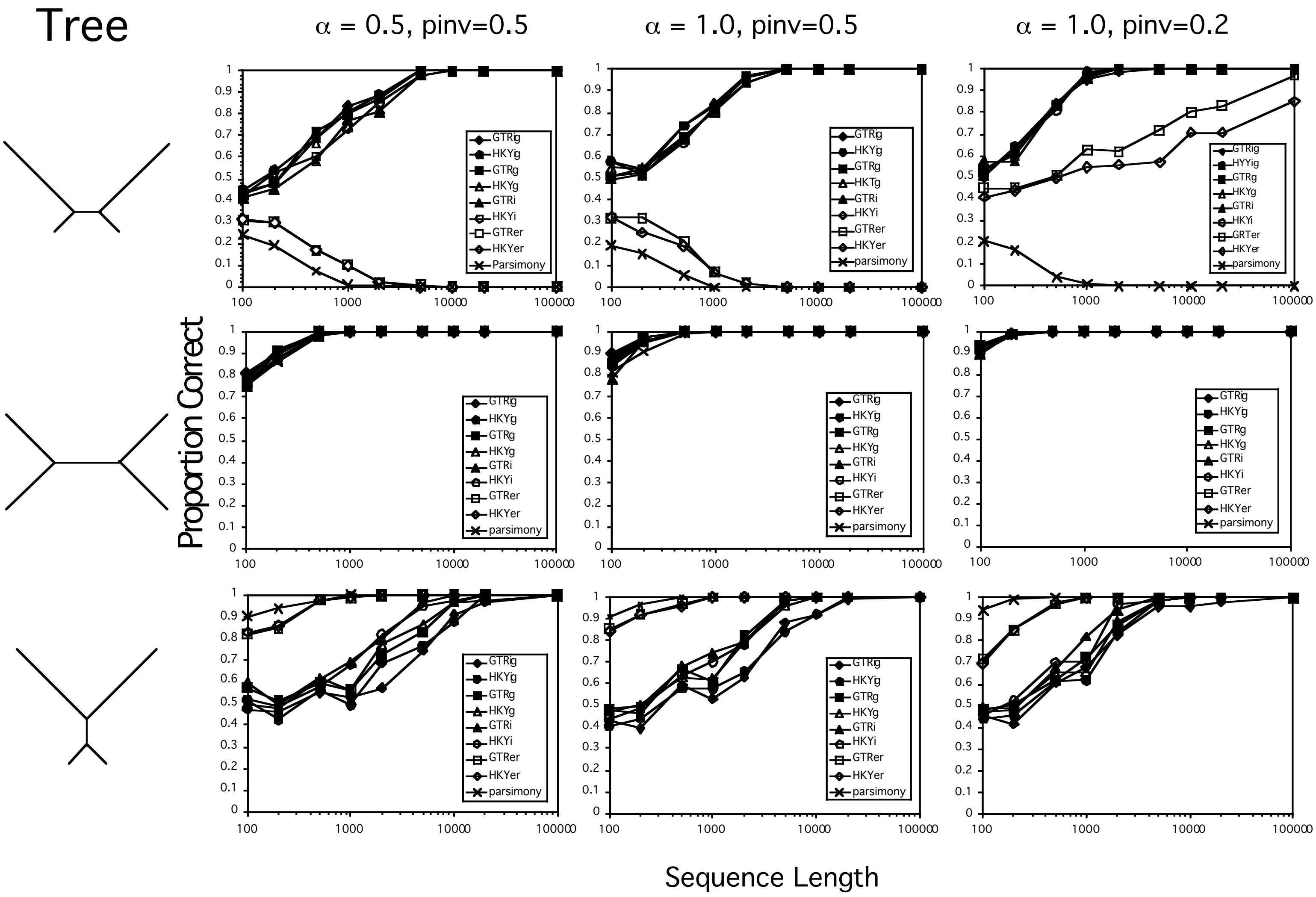
- Proportion of invariable sites
  - Some sites extremely unlikely to change due to strong functional or structural constraint (Hasegawa et al., 1985)
- Gamma-distributed rates
  - Rate variation assumed to follow a gamma distribution with shape parameter α
- Site-specific rates (another way to model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

Modeling ASRV with gamma distribution



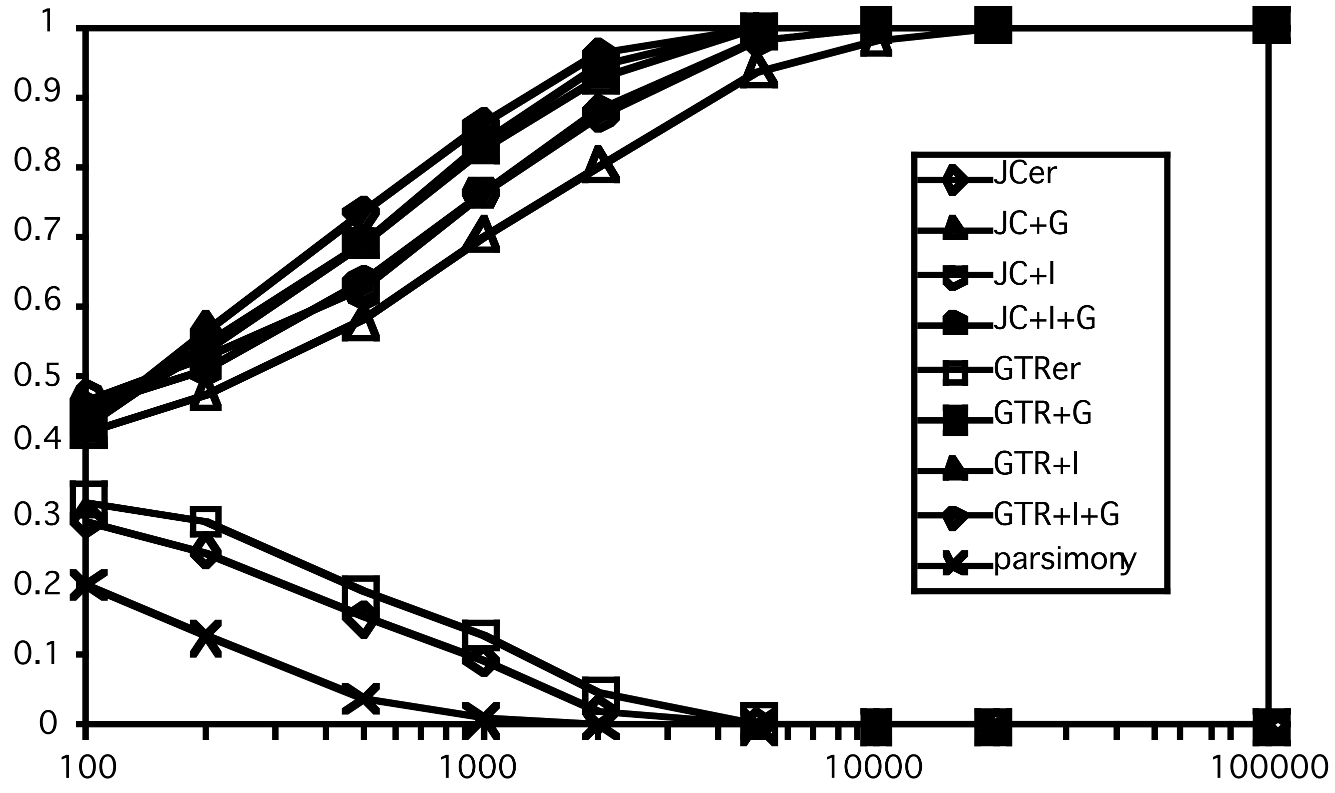
…can also include a proportion of “invariable” sites (p_{inv})

Performance of ML when its model is violated

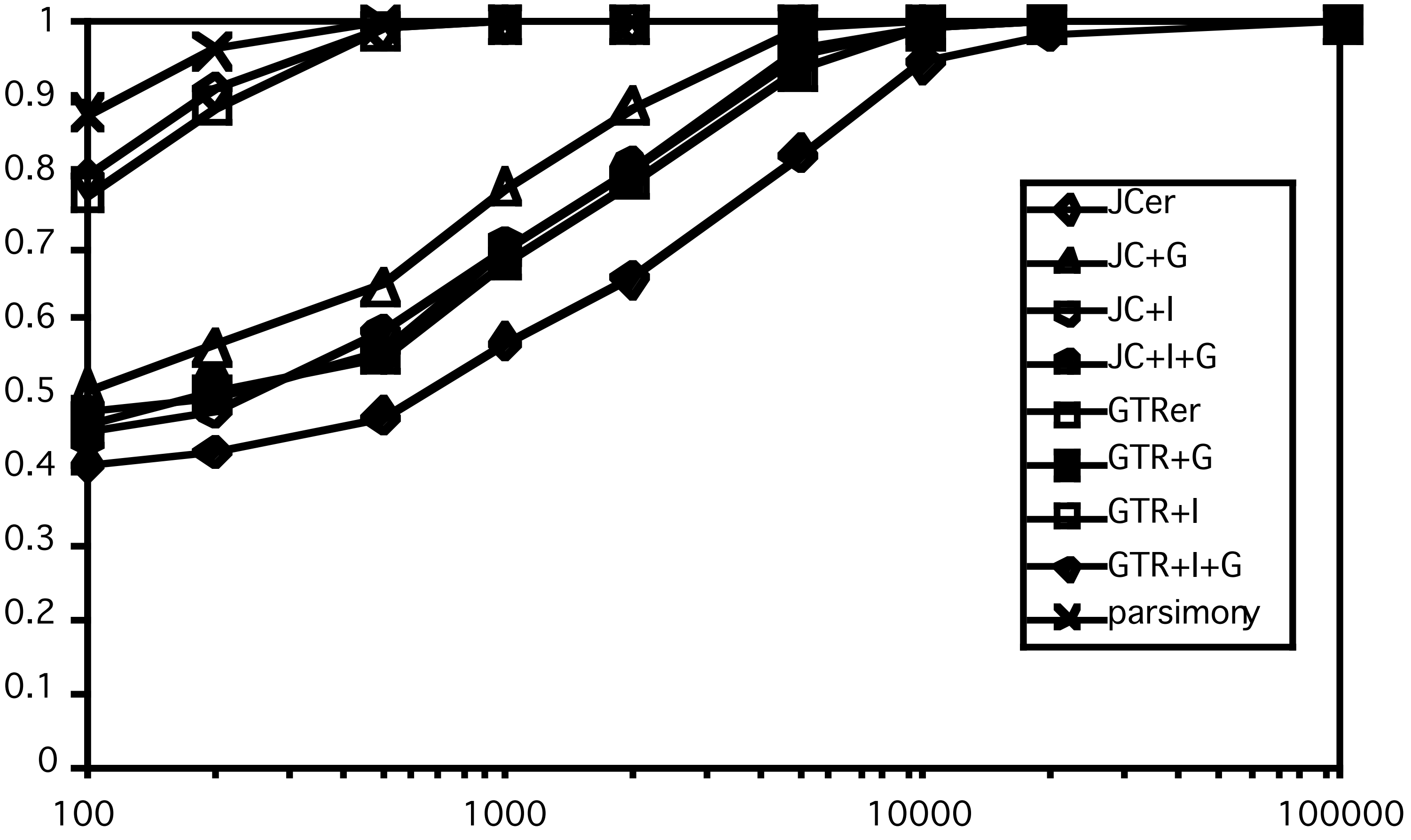


“MODERATE”–Felsenstein zone

α = 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 χ^2 with degrees-of-freedom equal to difference in number of free parameters

- Akaike information criterion (AIC)

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

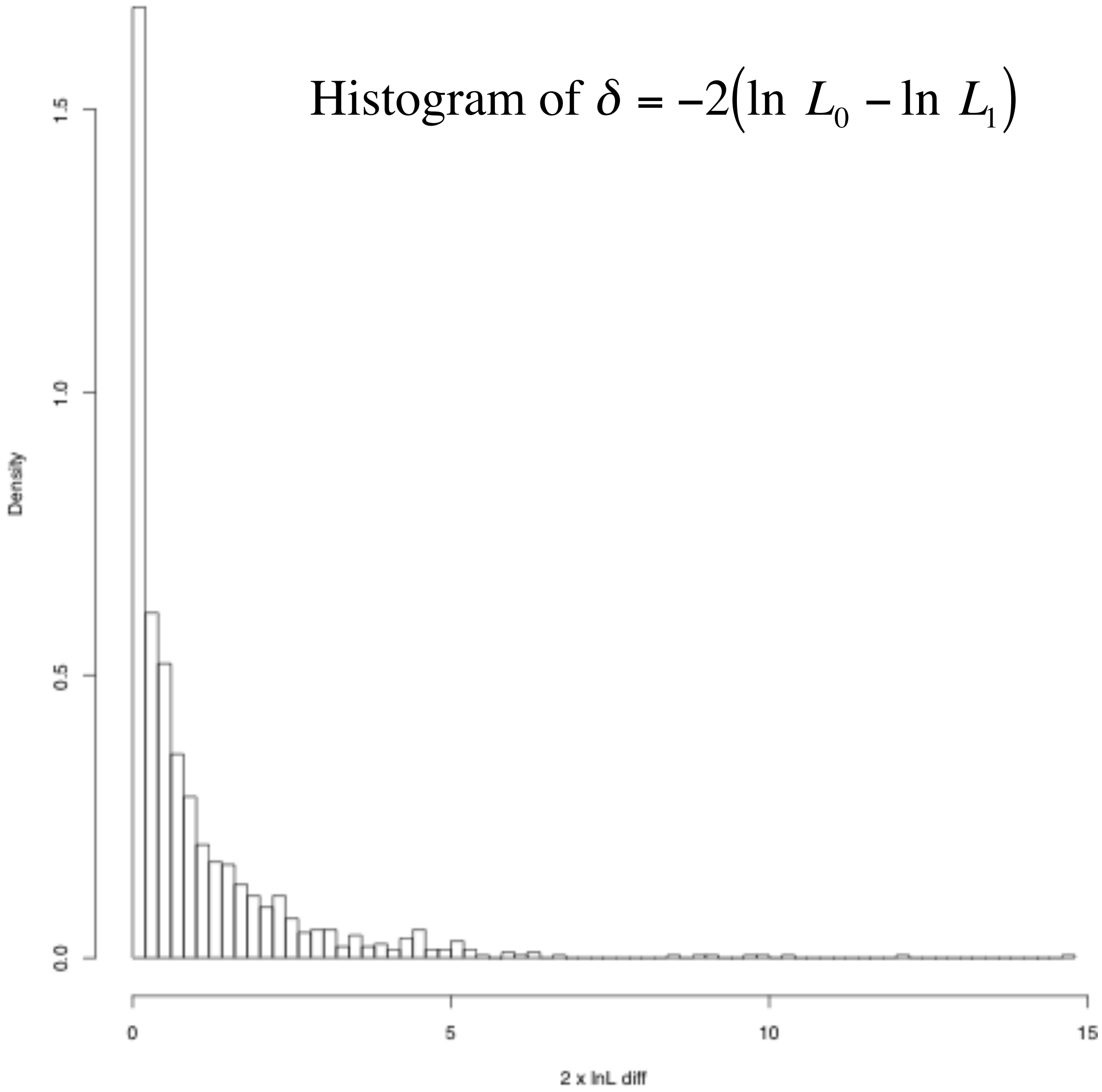
where *K* is the number of free parameters estimated

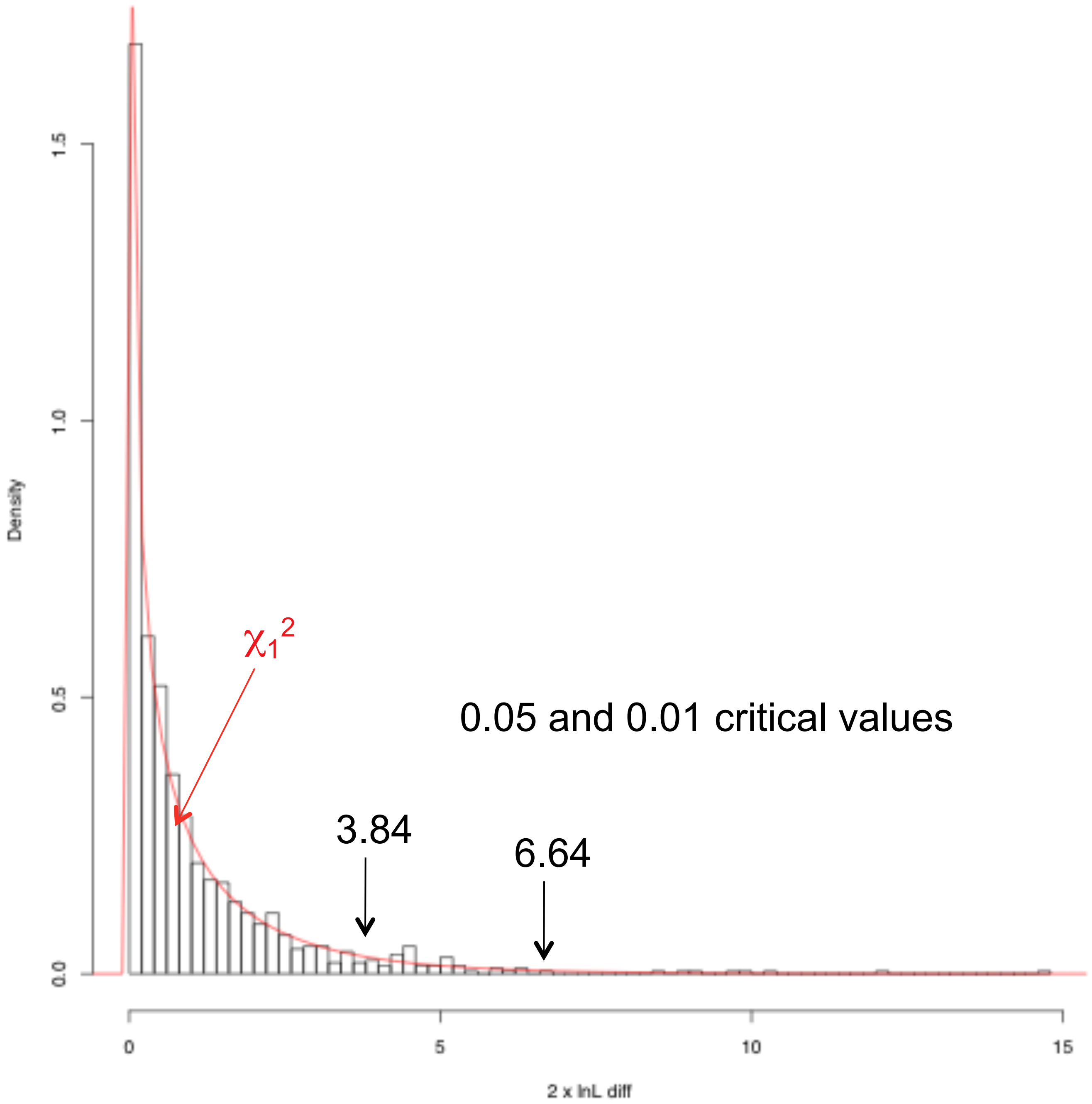
- Bayesian information criterion (BIC)

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

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

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




χ_1^2
0.05 and 0.01 critical values
3.84
6.64
Density
0.0
0.5
1.0
1.5
0
5
10
15
2 x lnL diff

What is PAUP*?

A multipurpose program for phylogenetic analysis

- Simple, intuitive interface
- Wide variety of analyses available in a single program (facilitates exploration)

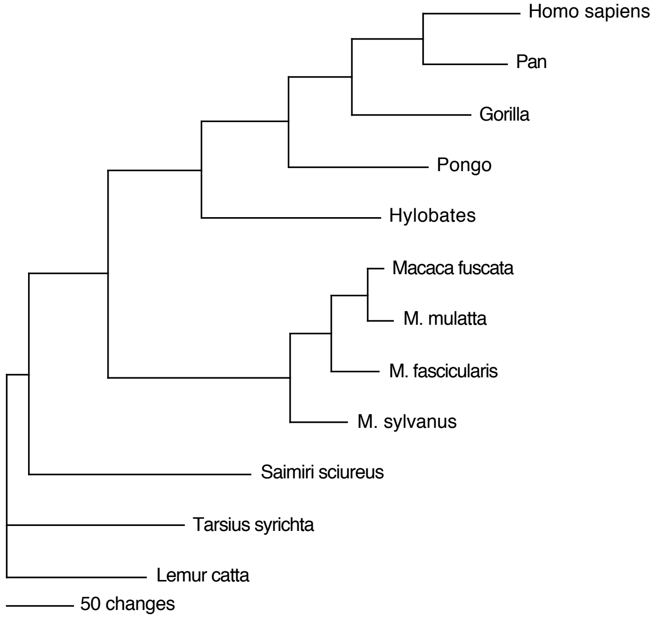
Searching for optimal evolutionary trees:

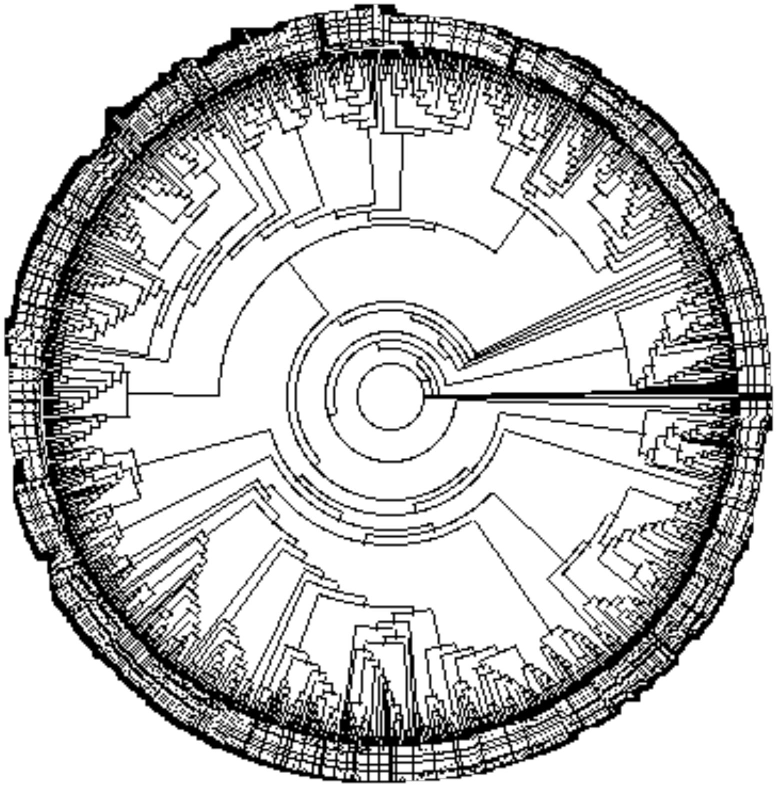
- Parsimony
- Distance
    - Minimum evolution
    - Least-squares
    - Weighted least squares (Fitch-Margoliash)
- Maximum likelihood (under a variety of models)

Tree-search algorithms

- Exact
  - Branch and bound
  - Exhaustive search
- Heuristic (local search)
  - Nearest-neighbor interchange (NNI)
  - Subtree pruning-regrafting (SPR)
  - Tree bisection-reconnection (TBR)
  - Edge contract-refine (2-ECR)
  - Star decomposition
  - Quartet puzzling
- Clustering/algorithmic
  - Neighbor-joining (including BioNJ)
  - UPGMA

Decent graphics





Confidence and hypothesis testing

- Bootstrap
- Jackknife
- Kishino-Hasegawa test
- Shimodaira-Hasegawa test
- Shimodaira approximately unbiased test
- Nonparametric Templeton and winning-sites tests
- Permutation tests
- Partition homogeneity (ILD) test

Models

- **DNA substition models** (both for distance and ML)
  - Jukes-Cantor
  - Kimura 2-parameter and 3ST
  - HKY85 and Felsenstein84
  - General time reversible (including any arbitrary submodel)
- **Amino acid models** (new)
  - PAM
  - JTT
  - mtREV
  - WAG
  - Any user-specified rate matrix
  - GTR

- **Among-site rate variation**
  - Gamma-distributed
  - Proportion of invariable sites
  - Gamma + P_{inv}
  - Site-specific

Other analyses and functions

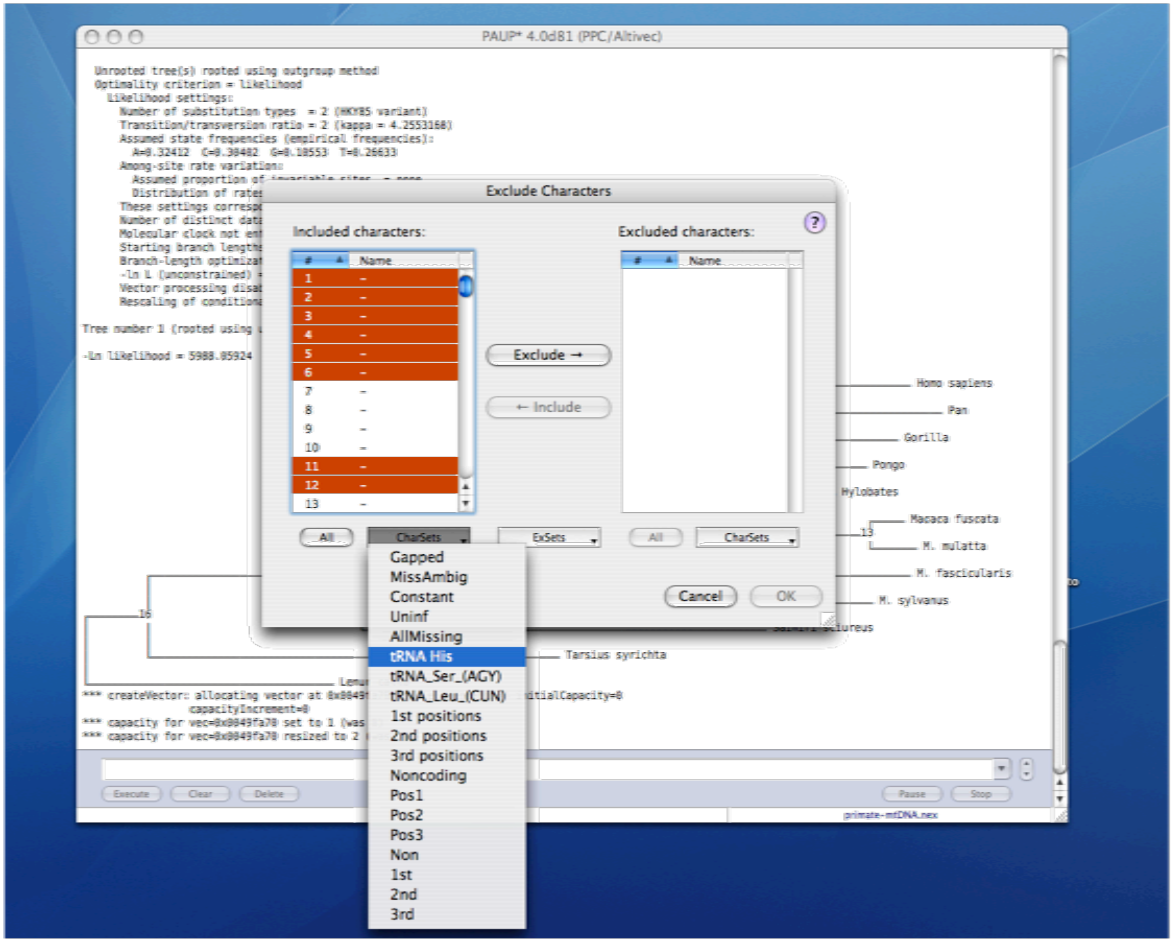
- **Summarizing agreement among trees**
  - Strict consensus
  - Majority-rule consensus
  - Adams consensus
  - Agreement subtrees
- **Filtering/sorting trees**
  - By compatibility with consensus
  - By tree score
- **Tree output and description**
  - Cladograms
  - Phylograms
  - Unrooted trees
- **Reconstruction of ancestral character states**
  - Parsimony
  - ML
- **Tree-to-tree distances** (RF, agreement metric, “ABC”)
- **Import/export of foreign formats** (PHYLIP, Mega, NBRF, Hennig/Nona/TNT)

Some new features

- Amino acid models
- Vectorized parsimony and likelihood calculations (Altivec, SSE)
- Multithreaded (pthreads, OpenMP) for multiprocessor and multicore machines
- Supertrees (MRP, strict consensus merger, others?)
- Soon...
    - Simple checkpointing
    - Parallel tree evaluation (MPI and PVM)
    - Improved tree-search heuristics (e.g., “ratchet”)

PAUP* 4.0 Platforms

OS X (native)

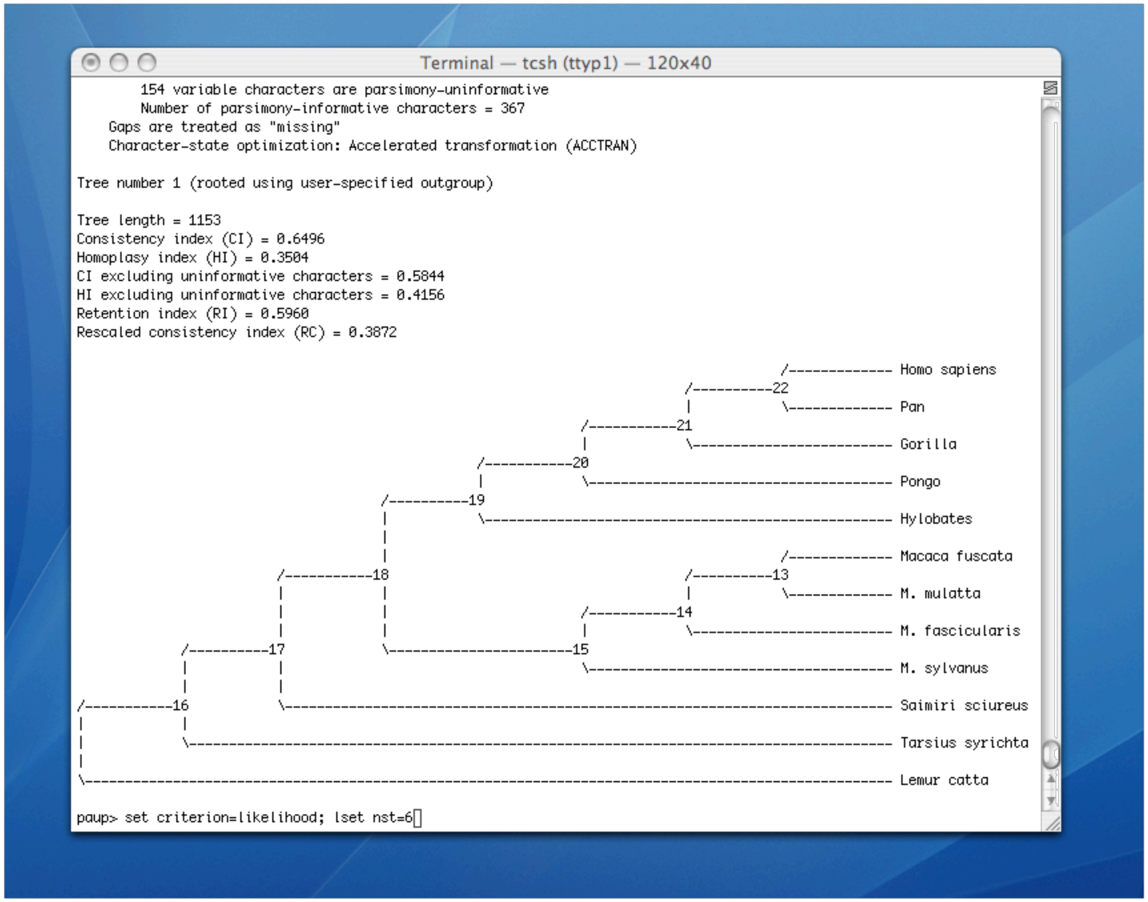


```
Unrooted tree(s) rooted using outgroup method
Optimality criterion = likelihood
  Likelihood settings:
    Number of substitution types  = 2 (HKY85 variant)
    Transition/transversion ratio = 2 (kappa = 4.2553168)
    Assumed state frequencies (empirical frequencies):
      A=0.32412  C=0.30482  G=0.10553  T=0.26633
    Among-site rate variation:
      Assumed proportion of invariable sites  = none
      Distribution of rates
    These settings corresp
    Number of distinct data
    Molecular clock not enf
    Starting branch lengths
    Branch-length optimizat
    -ln L (unconstrained) =
    Vector processing disab
    Rescaling of conditiona

Tree number 1 (rooted using

-ln likelihood = 5988.05924
```

```
*** createVector: allocating vector at 0x0049                    itialCapacity=0
                  capacityIncrement=0
*** capacity for vec=0x0049fa70 set to 1 (was
*** capacity for vec=0x0049fa70 resized to 2
```

PAUP* 4.0 Platforms

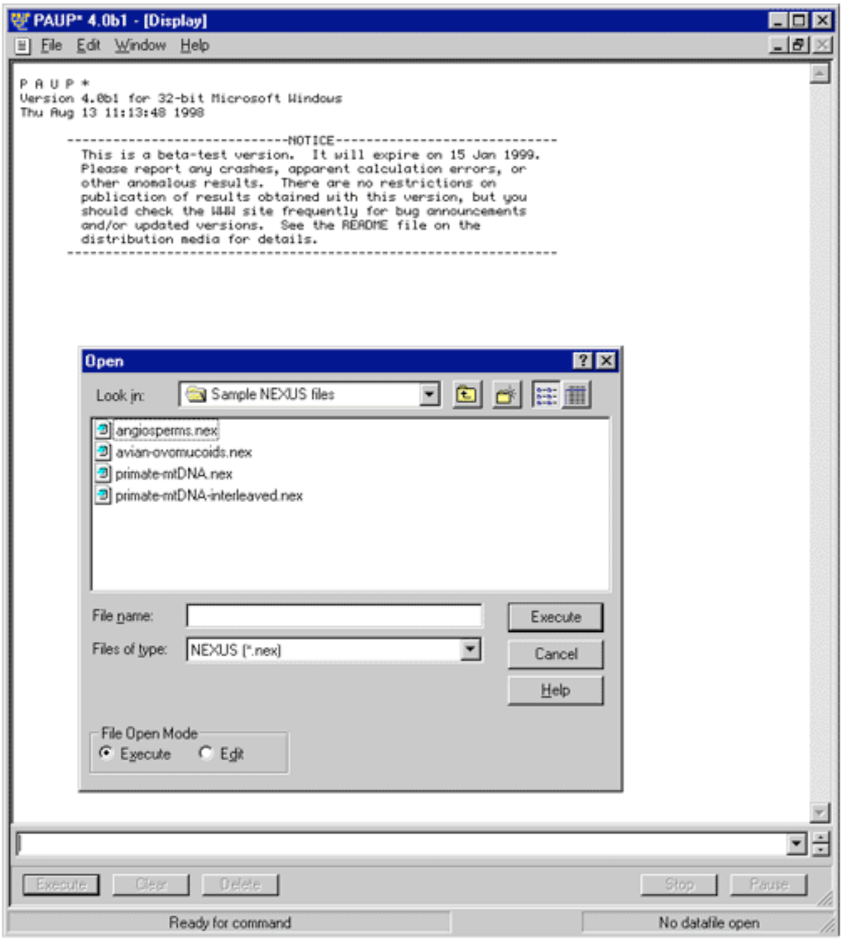
Linux/Unix/OS X Terminal



```
        154 variable characters are parsimony-uninformative
        Number of parsimony-informative characters = 367
    Gaps are treated as "missing"
    Character-state optimization: Accelerated transformation (ACCTRAN)

Tree number 1 (rooted using user-specified outgroup)

Tree length = 1153
Consistency index (CI) = 0.6496
Homoplasy index (HI) = 0.3504
CI excluding uninformative characters = 0.5844
HI excluding uninformative characters = 0.4156
Retention index (RI) = 0.5960
Rescaled consistency index (RC) = 0.3872

                                                                                        /------------- Homo sapiens
                                                                            /----------22
                                                                            |           \------------- Pan
                                                               /-----------21
                                                               |            \------------------------- Gorilla
                                                  /-----------20
                                                  |            \-------------------------------------- Pongo
                                      /----------19
                                      |           \--------------------------------------------------- Hylobates
                                      |
                                      |                                                 /------------- Macaca fuscata
                         /-----------18                                     /----------13
                         |            |                                     |           \------------- M. mulatta
                         |            |                        /-----------14
                         |            |                        |            \------------------------- M. fascicularis
             /----------17            \-----------------------15
             |           |                                     \-------------------------------------- M. sylvanus
             |           |
/-----------16           \---------------------------------------------------------------------------- Saimiri sciureus
|            |
|            \---------------------------------------------------------------------------------------- Tarsius syrichta
|
\----------------------------------------------------------------------------------------------------- Lemur catta

paup> set criterion=likelihood; lset nst=6
```

PAUP* 4.0 Platforms

Microsoft Windows (full GUI almost finished)



```
P A U P *
Version 4.0b1 for 32-bit Microsoft Windows
Thu Aug 13 11:13:48 1998

      ----------------------------NOTICE----------------------------
        This is a beta-test version.  It will expire on 15 Jan 1999.
        Please report any crashes, apparent calculation errors, or
        other anomalous results.  There are no restrictions on
        publication of results obtained with this version, but you
        should check the WWW site frequently for bug announcements
        and/or updated versions.  See the README file on the
        distribution media for details.
      ---------------------------------------------------------------
```

Look in: Sample NEXUS files

- angiosperms.nex
- avian-ovomucoids.nex
- primate-mtDNA.nex
- primate-mtDNA-interleaved.nex

File name:

Files of type: NEXUS (*.nex)

File Open Mode

● Execute ○ Edit

Ready for command

No datafile open

Helpers/Collaborators

Jim Wilgenbusch (support, documentation, production)

Chuck Bell (coauthor of version 4.0 manual)

John Huelsenbeck

Paul Lewi
s

David Bryant

Peter Waddell