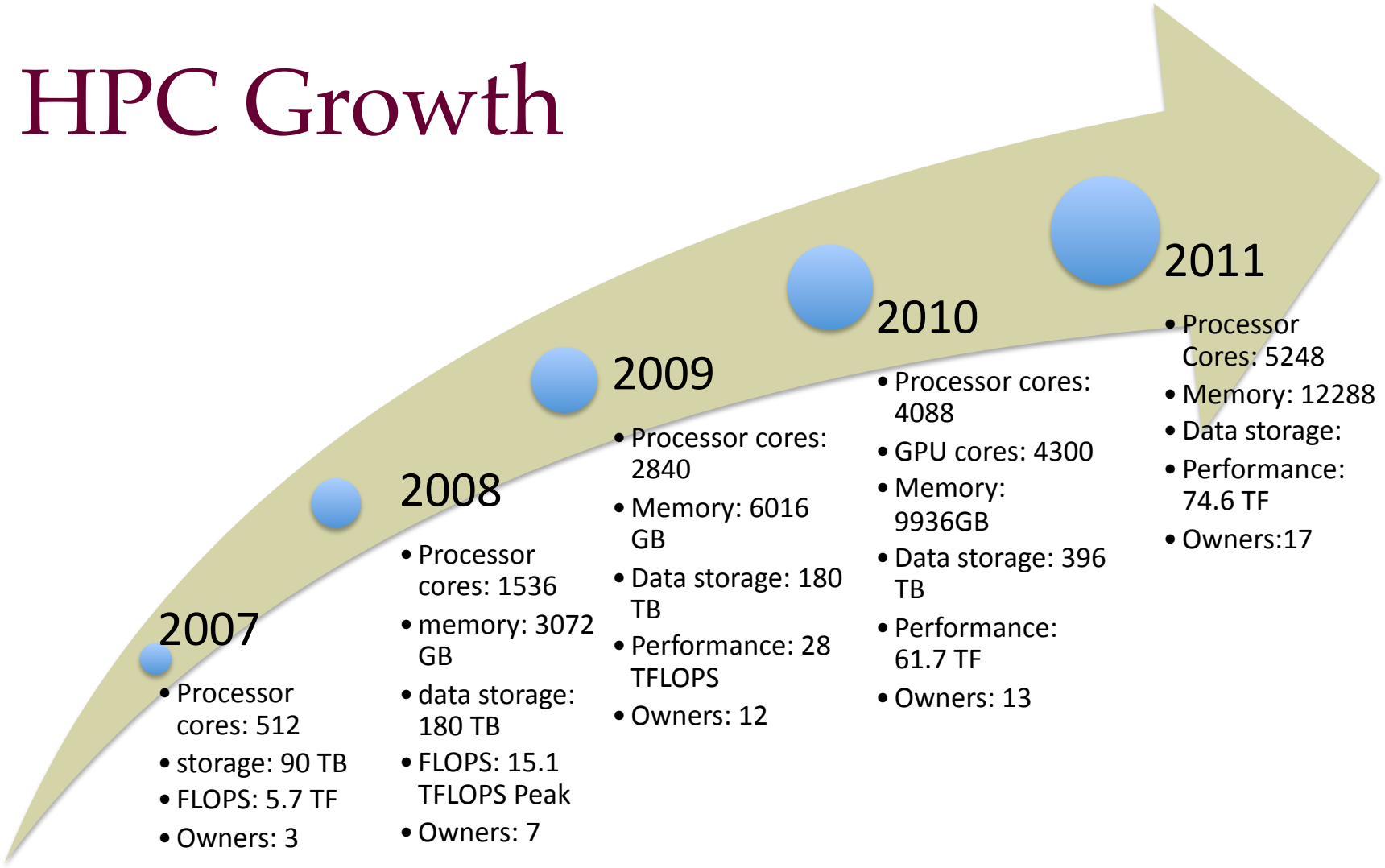
Phylogenetic Inference

Jim Wilgenbusch

Department of Scientific Computing
Florida State University

Florida State University

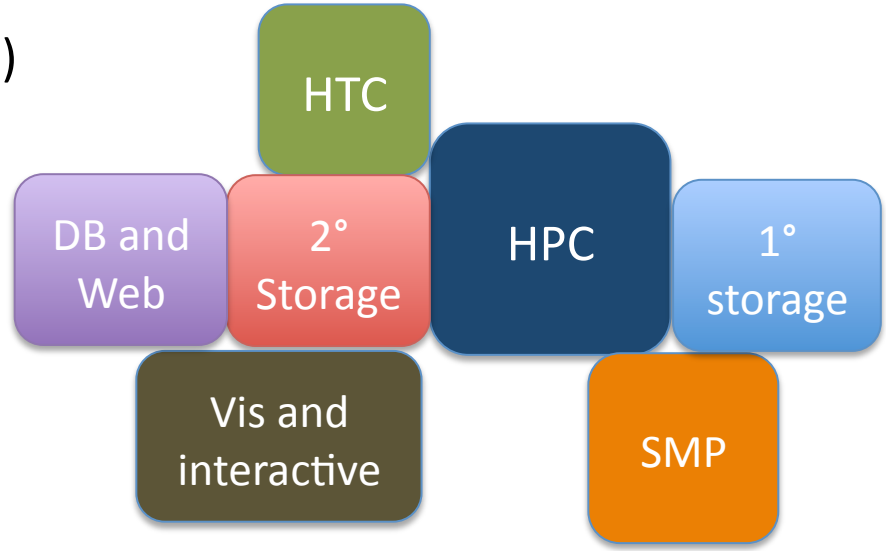
HPC Growth


2007
• Processor cores: 512
• storage: 90 TB
• FLOPS: 5.7 TF
• Owners: 3
2008
• Processor cores: 1536
• memory: 3072 GB
• data storage: 180 TB
• FLOPS: 15.1 TFLOPS Peak
• Owners: 7
2009
• Processor cores: 2840
• Memory: 6016 GB
• Data storage: 180 TB
• Performance: 28 TFLOPS
• Owners: 12
2010
• Processor cores: 4088
• GPU cores: 4300
• Memory: 9936GB
• Data storage: 396 TB
• Performance: 61.7 TF
• Owners: 13
2011
• Processor Cores: 5248
• Memory: 12288
• Data storage:
• Performance: 74.6 TF
• Owners:17

Support Multidisciplinary Research

•Diverse Research Cyber-Infrastructure

- One size does NOT fit all
- High Performance Computing (HPC)
  - Distributed Memory (MPI framework)
  - Shared Memory (OpenMP framework)
- High Throughput Computing (HTC)
- Diverse storage requirements
- Web portals and scientific databases
- Remote Visualization (Vis) and Interactive Computing



Broad application support

- Macromolecules
- Ground Water
- Genetics
- Physics Procedia
- Bioinformatics
- Systematic Biology
- Journal of Biogeography
- Journal of Applied Remote Sensing
- Journal of Chemical Theory and Computation
- Physical Review Letters
- Journal of Physical Chemistry
- Journal of Physics: Condensed Matter
- Proceeding of the National Academy of Science
- Biophysical Journal
- Journal Chemical Theory Computation
- International Journal of Human Modeling and Simulation
- International Journal of Crashworthiness
- Journal: J. Phys. Chem.
- PLoS Pathogens
- Journal of Virology
- Journal of the American Chemical Society
- The Journal of Chemical Physics
- PLoS Biology
- Ocean Modeling
- Journal of Computer-Aided Molecular Design
- Journal of Geophysical Research – Atmospheres
- Evolution
- Atmosphere-Ocean
- Ecological Complexity
- Water Resources Research
- International Journal of Impact Engineering
- SPE Reservoir Evaluation & Engineering
- Journal of Biogeography
- Political Analysis
- … and many more

Survival in the Academic Jungle

Darwinian evolution meets high performance computing

James Wilgenbusch

My background is in evolutionary biology, so I frequently find myself borrowing concepts like "bet-hedging," "hybrid vigor" and "punctuated equilibrium" to help make decisions related to acquiring and supporting high performance computing assets in a highly variable and often unpredictable academic environment. Borrowing from this field is not new. Economics has a long history of using evolutionary theory to explain and make predictions regarding the behavior of complex systems. Therefore, it should come as no surprise that concepts designed originally to explain organic phenomena also have value in the metal and silicon world of scientific computing support. Despite their distinctly different appearances, at some level, the forces driving these systems are very similar.

I could (and I evolutionary perspective when acquiring and supporting research computing assets is that it teaches you to carefully and honestly consider

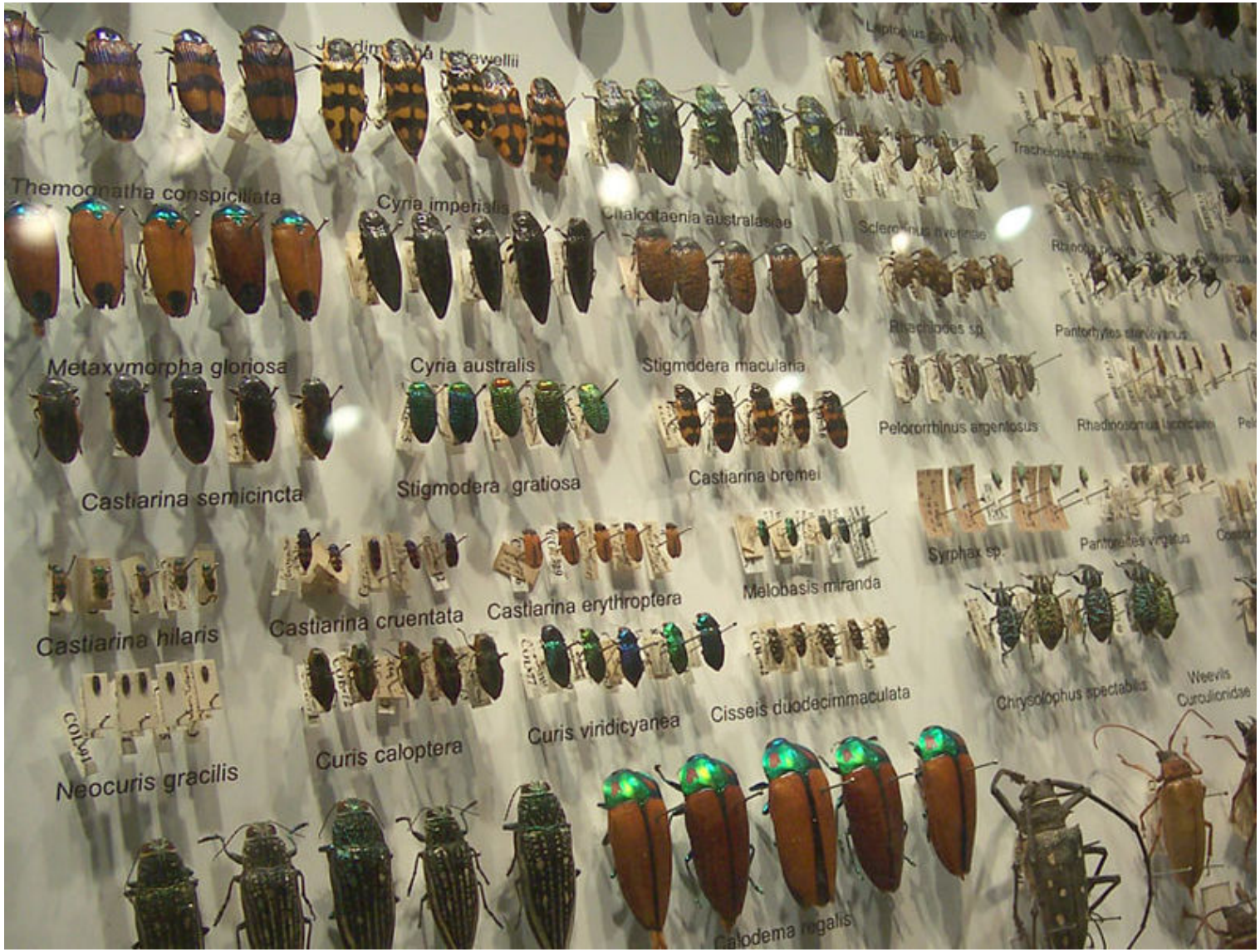
- what is being optimized
- which of the many variables that you need to evaluate are predictable
- how the previous two considerations will change over time.

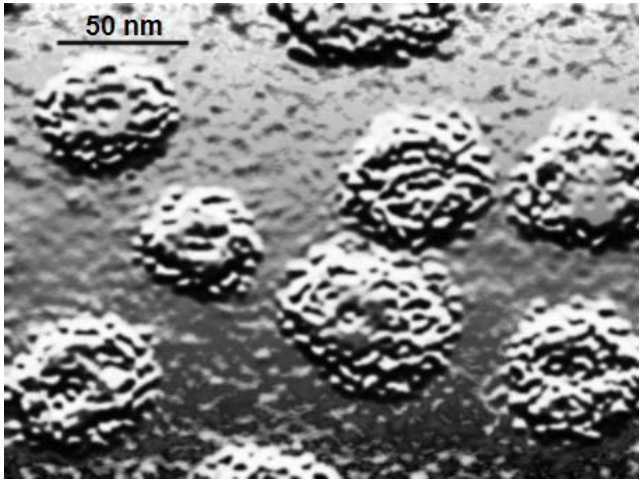
If this sounds like a hard problem, that is because it is.

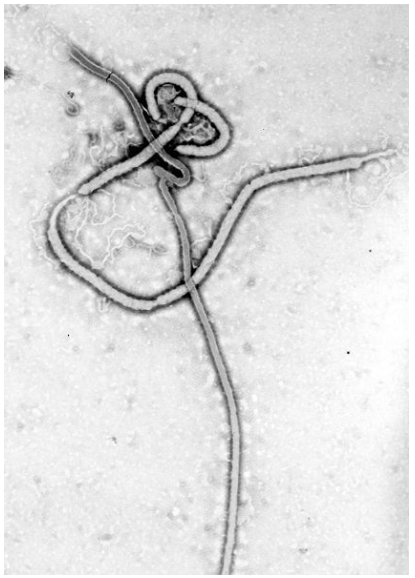
"It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change."
— Charles Darwin

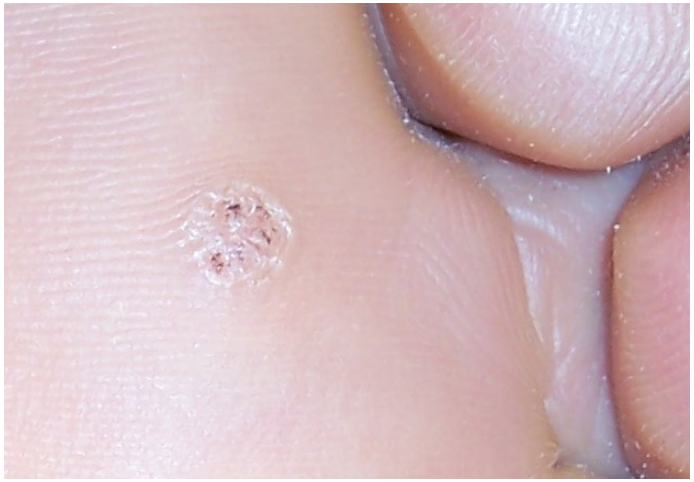
OPTIMIZE WHAT?

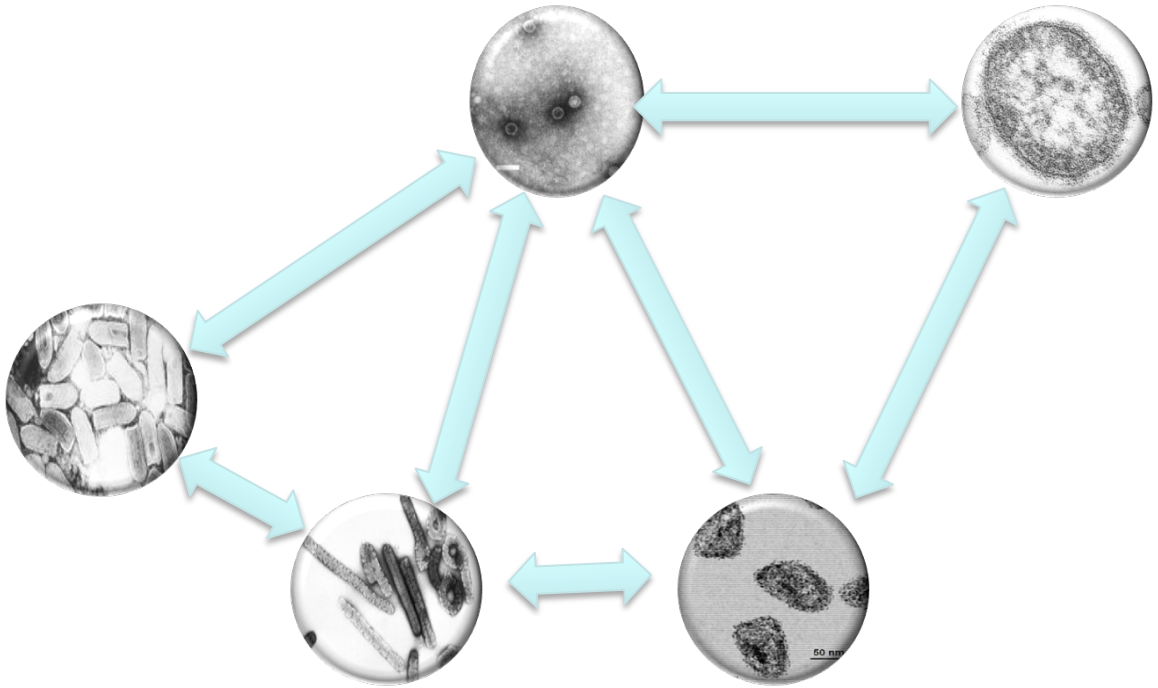
Five years ago, Florida State University (FSU) faced a dilemma not uncommon to large universities supporting diverse research programs. The problem was that researchers were complaining that they lacked adequate computing and storage resources to support their research programs and, by extension, to fulfill their obligations to external funding agencies. The dilemma was that these complaints were made while, at the same time, FSU was supporting a shared supercomputer that had recently run a

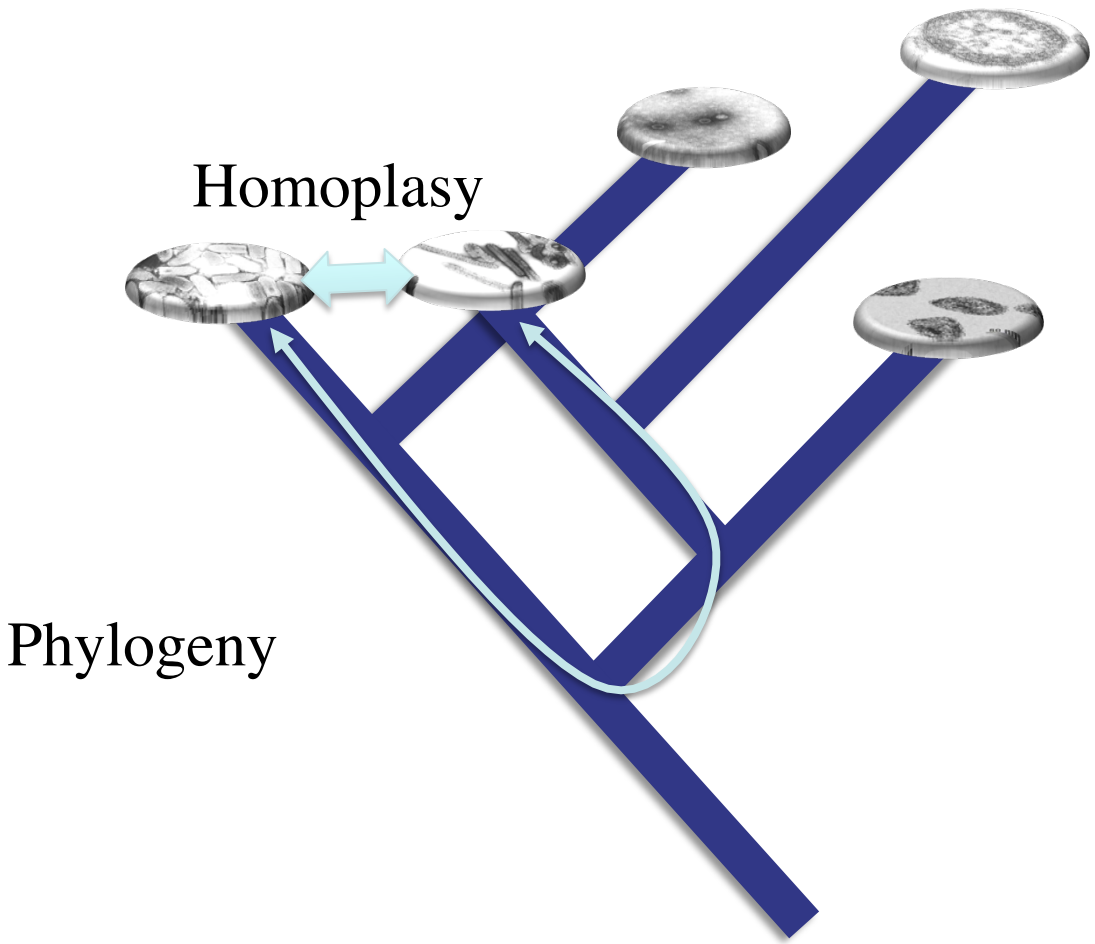

Themognatha conspicillata
Cyria imperialis
Chalcotaenia australasiae
Metaxymorpha gloriosa
Cyria australis
Stigmodera macularia
Castiarina semicincta
Stigmodera gratiosa
Castiarina bremei
Castiarina hilaris
Castiarina cruentata
Castiarina erythroptera
Melobasis miranda
Syrphax sp.
Neocuris gracilis
Curis caloptera
Curis viridicyanea
Cisseis duodecimmaculata
Chrysolophus spectabilis
Weevils
Curculionidae
Calodema regalis


50 nm






50 nm


Homoplasy
Phylogeny

Evolutionary Biology

- Evolutionary thinking
- What are the products of evolution?
- What motivates questions in evolutionary biology?
- Some applications

Nature, it seems, is the popular name
For milliards and milliards and milliards
Of particles playing their infinite game
Of billiards and billiards and billiards.

Piet Hein

Evolutionary Biology

- Evolutionary thinking
- What are the products of evolution?
- What motivates questions in evolutionary biology?
- Some applications

Evolutionary Biology

- Evolutionary thinking
- What are the products of evolution?
- What motivates questions in evolutionary biology?
- Some applications

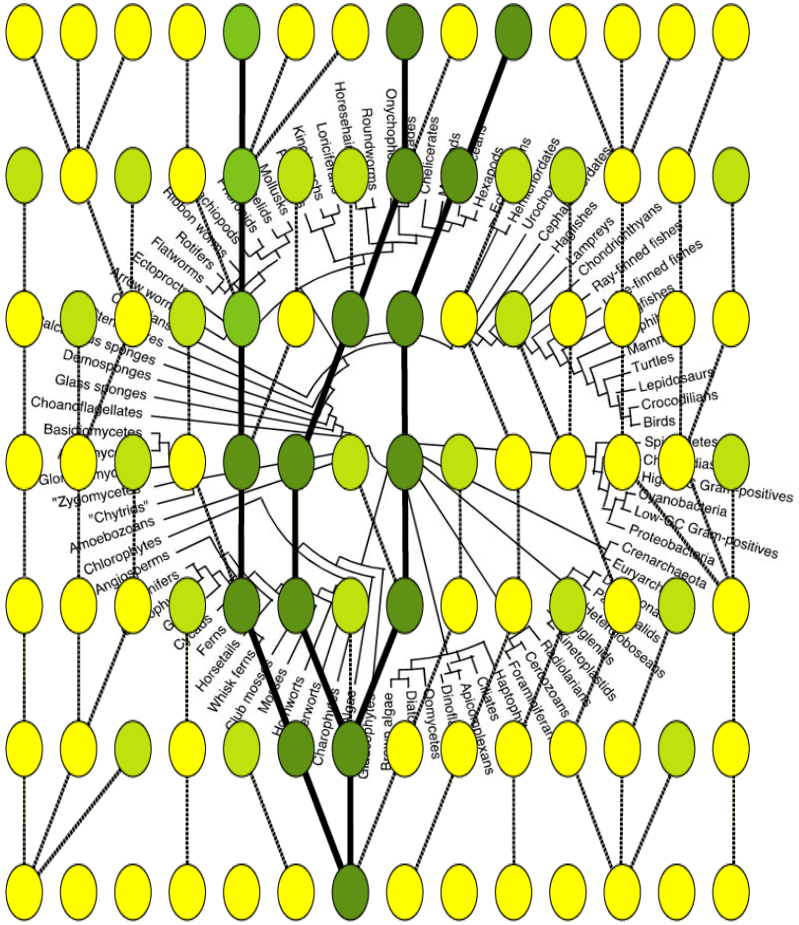
Evolutionary Biology

- Evolutionary thinking
- What are the products of evolution?
- What motivates questions in evolutionary biology?
- Some applications

- Systematics
- Conservation biology
  - Faith, 1992 (phylogeny and conservation priorities)
  - Baker and Palumbi, 1994 (illegal whale hunting)
- Epidemiology
  - Bush et al. 1999 (predictive evolution)
- Forensics
  - Ou, C. et al. 1992, Hillis and Huelsenbeck, 1994 (dental practice HIV transmission)
- Gene function prediction
  - Chang and Donoghue, 2000; Bader, et al., 2001
- Drug Development
  - Halbur, et al., 1994

Scale of Diversity in Time and Space

Modeling Perspective

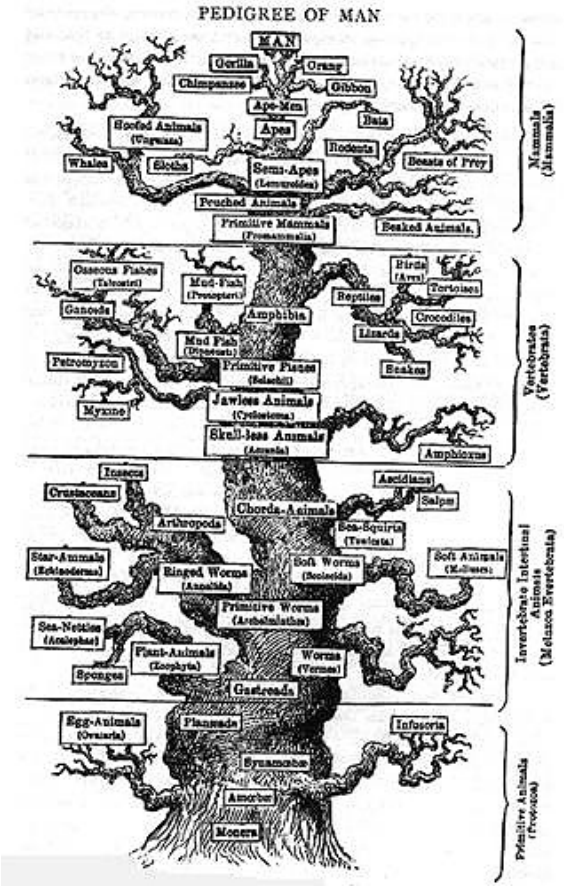


- Changes at or below the species level
  - Interested in changes in allele frequencies
  - Mutation
  - Selection (Natural and otherwise)
  - Genetic Drift
  - Gene flow or Migration
- Changes above species level
  - Compounded effects of microevolutionary processes
- Evolution is a unified theory
  - Models require the notion of scale

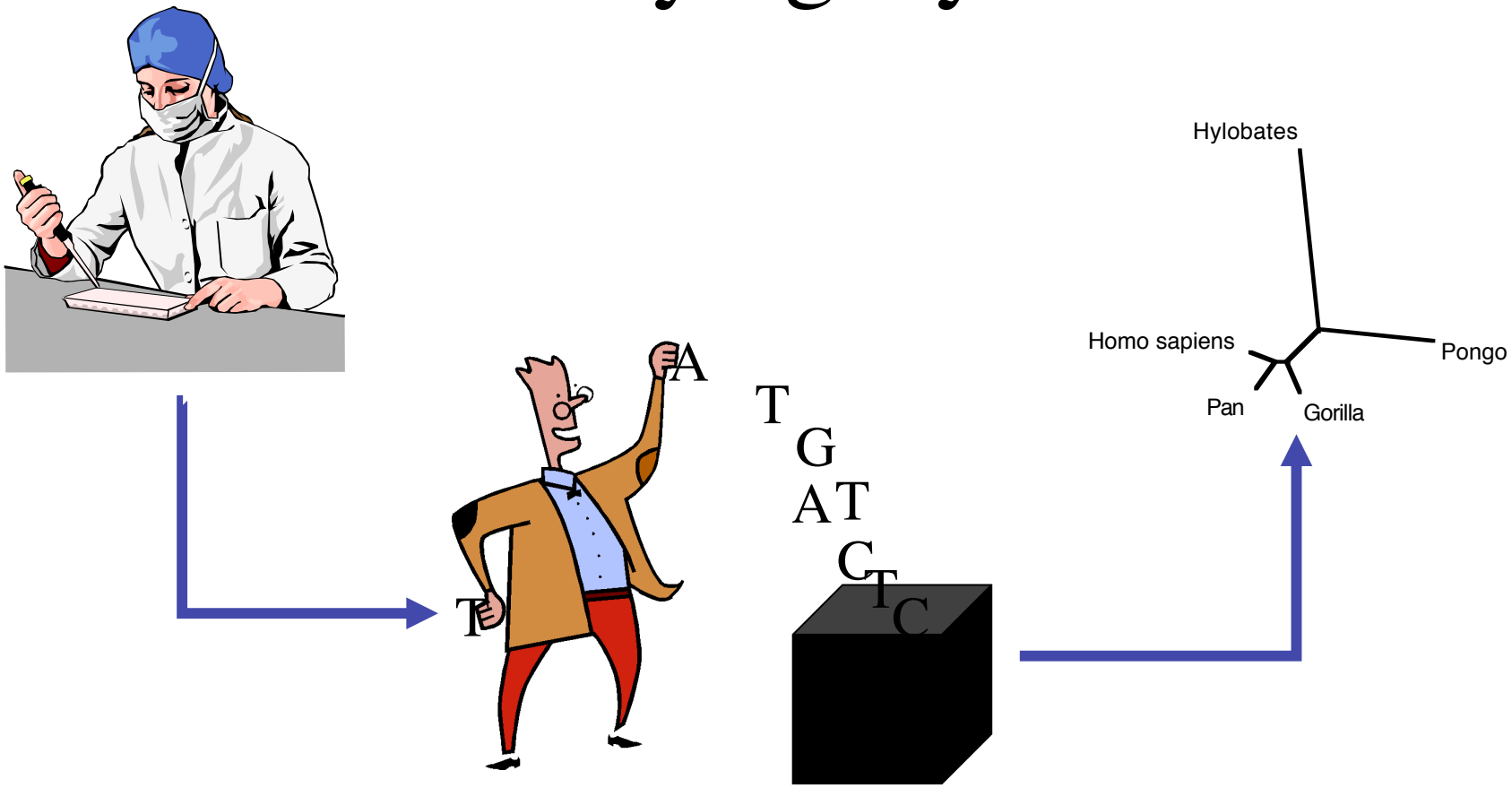
Macroevolution

Ernst Haeckel's "Geneological Tree", 1879

- Phylogenetic Framework
  - Involves an attempt to estimate the evolutionary history of a collection of organisms (taxa) or other biological objects.
- Fundamental part of modern biology
- Two major endeavors
  - estimating the evolutionary tree (branching order, branch lengths)
  - using the trees (phylogenies) as analytical framework for further evolutionary study



"Typical" procedure used to infer Phylogeny

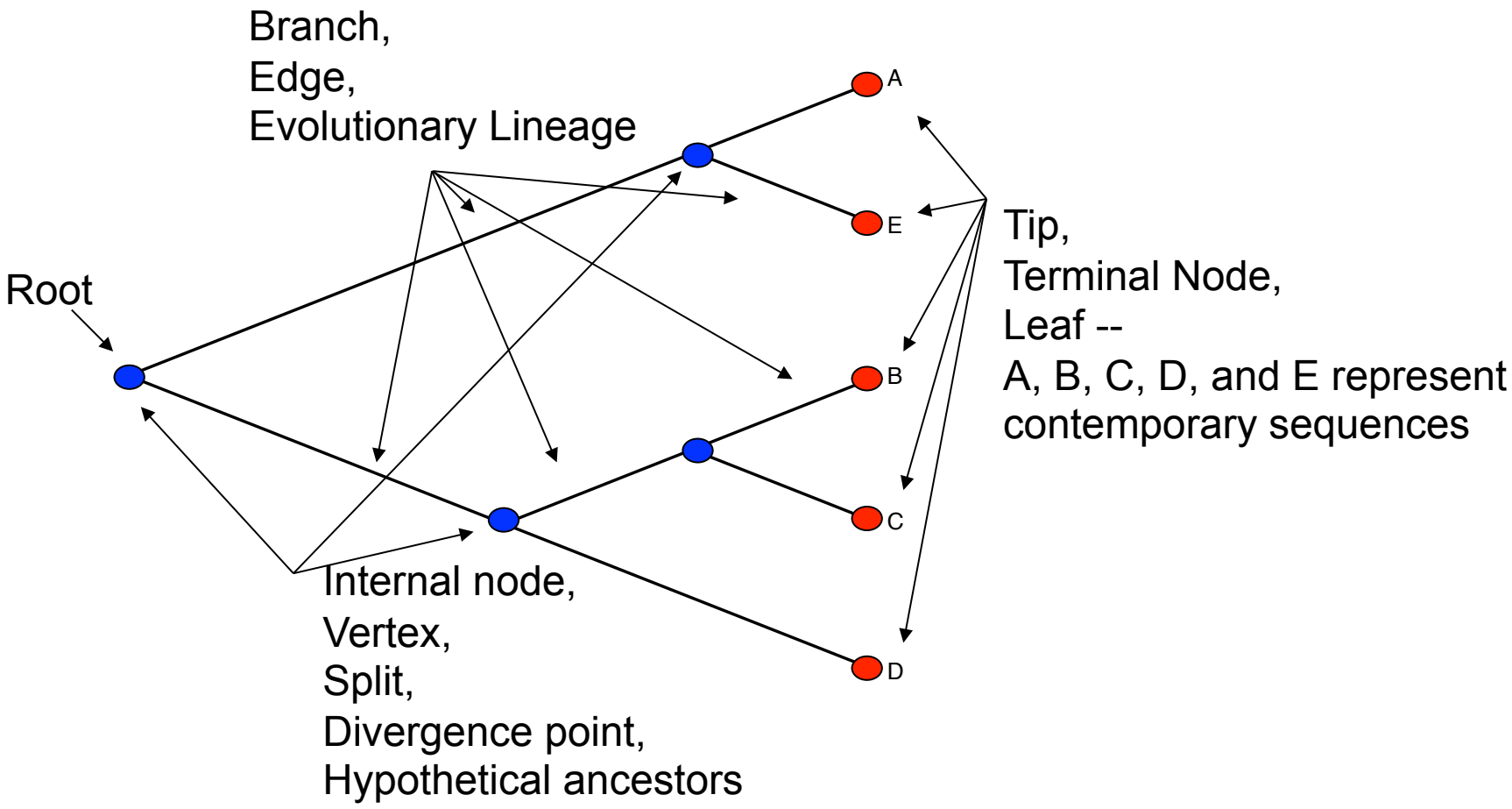


Phylogenetic Inference

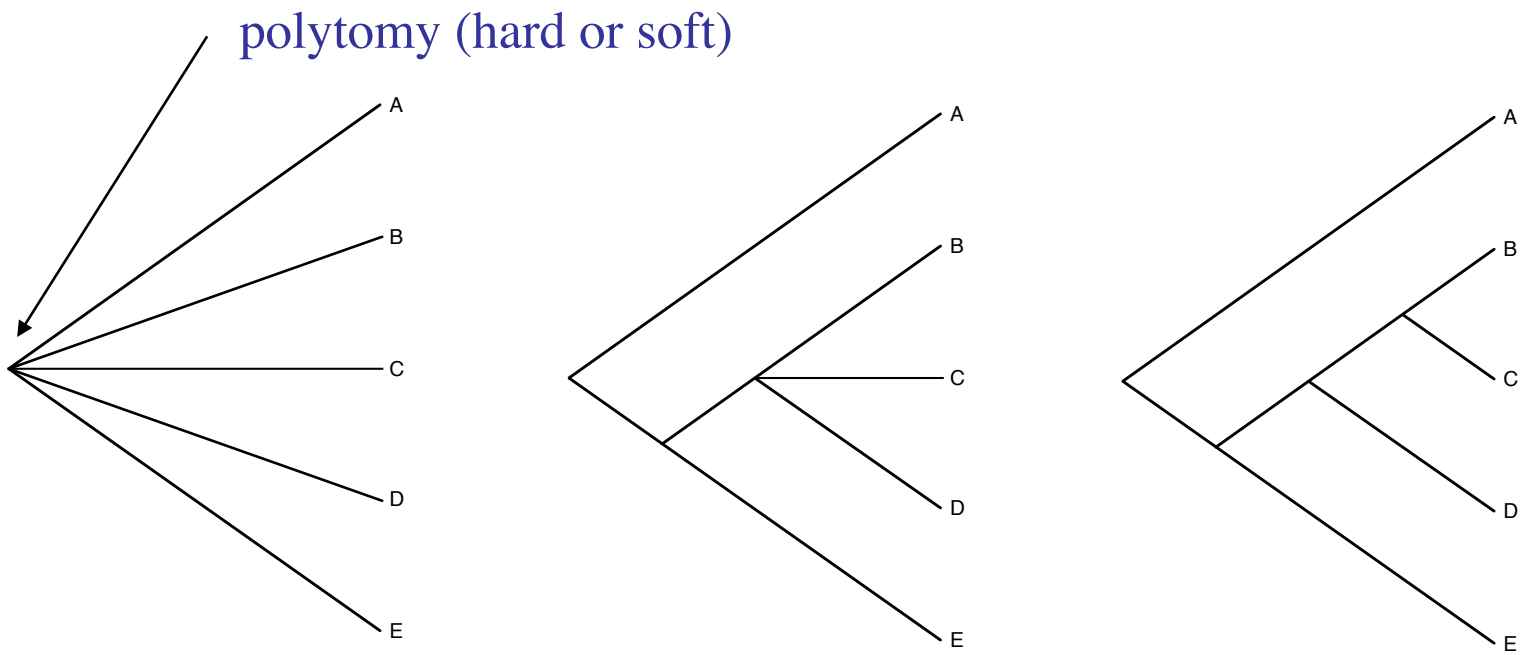
- Score the Phylogeny (small problem*)
  - Maximum parsimony
  - Maximum likelihood
  - Distance (Least-squares and Minimum Evolution)
- Search for the Best Phylogeny (large problem*)
  - Exact search
  - Heuristic search
  - Stochastic search
- Test the Reliability of Results
  - Support for individual tree nodes
  - Support for complete tree topologies

* From a computation perspective

Common Tree Terms



Common Tree Terms

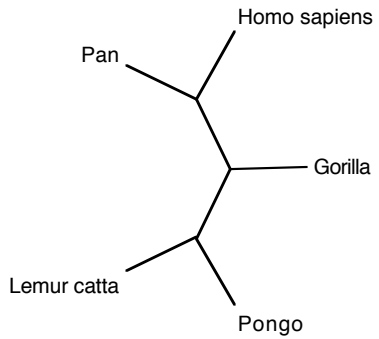


Star Tree
completely unresolved

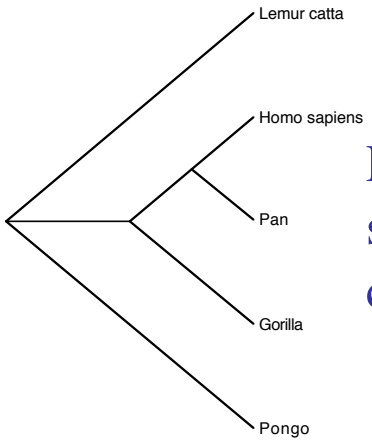
partially unresolved

Bifurcating Tree
completely resolved

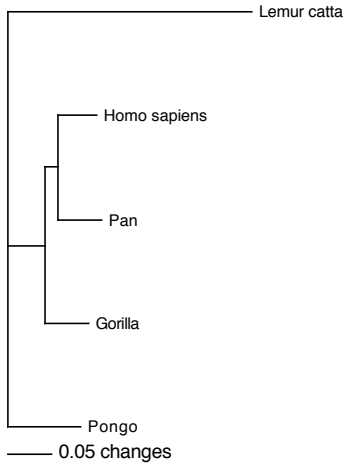
Common Tree Representations



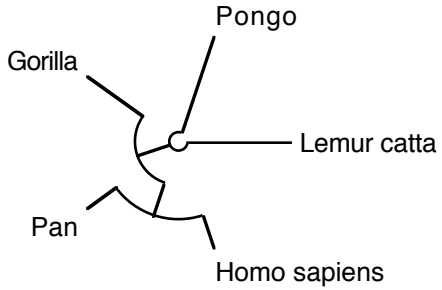
Unrooted cladogram



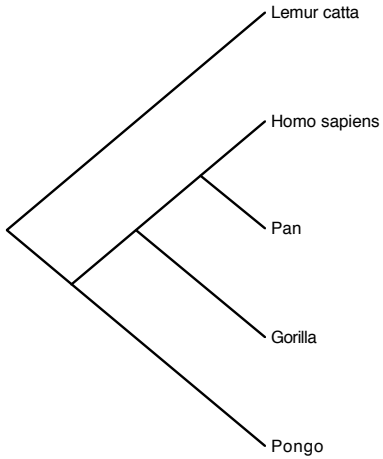
Rooted slanted cladogram



Rooted phylogram



Circle tree



Rooted slanted cladogram with ingroup monophyletic

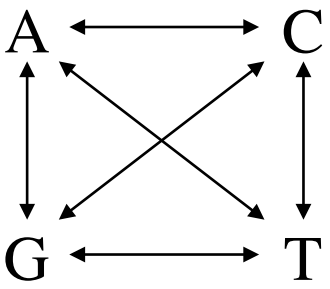
Parsimony (optimality criterion)

- In general, choose the tree requiring the fewest number of character-state changes (AKA, steps)
  - In other words, minimize the number of *ad hoc* assumptions (e.g., convergences, parallelisms, reversals).
- Homoplasy is typically used to mean convergences, parallelisms, and reversals
- Assume character independence; e.g., can calculate length required by each character and sum over characters to get total tree length

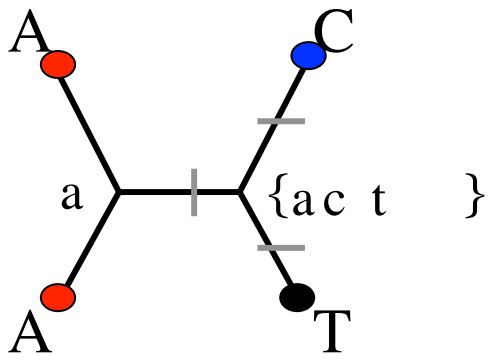
Parsimony (Small problem)

Minimize the number of *ad hoc* assumptions on a given tree (e.g., convergences, parallelisms, reversals)

Fitch parsimony (unordered/nonadditive): Each change counts 1 step, regardless of the nature of this change.



(1) …GGACAAGTTTA…
(2) …AGACAACTCTA…
(3) …GGATACGTTAA…
(4) …GGATATCCTAG…



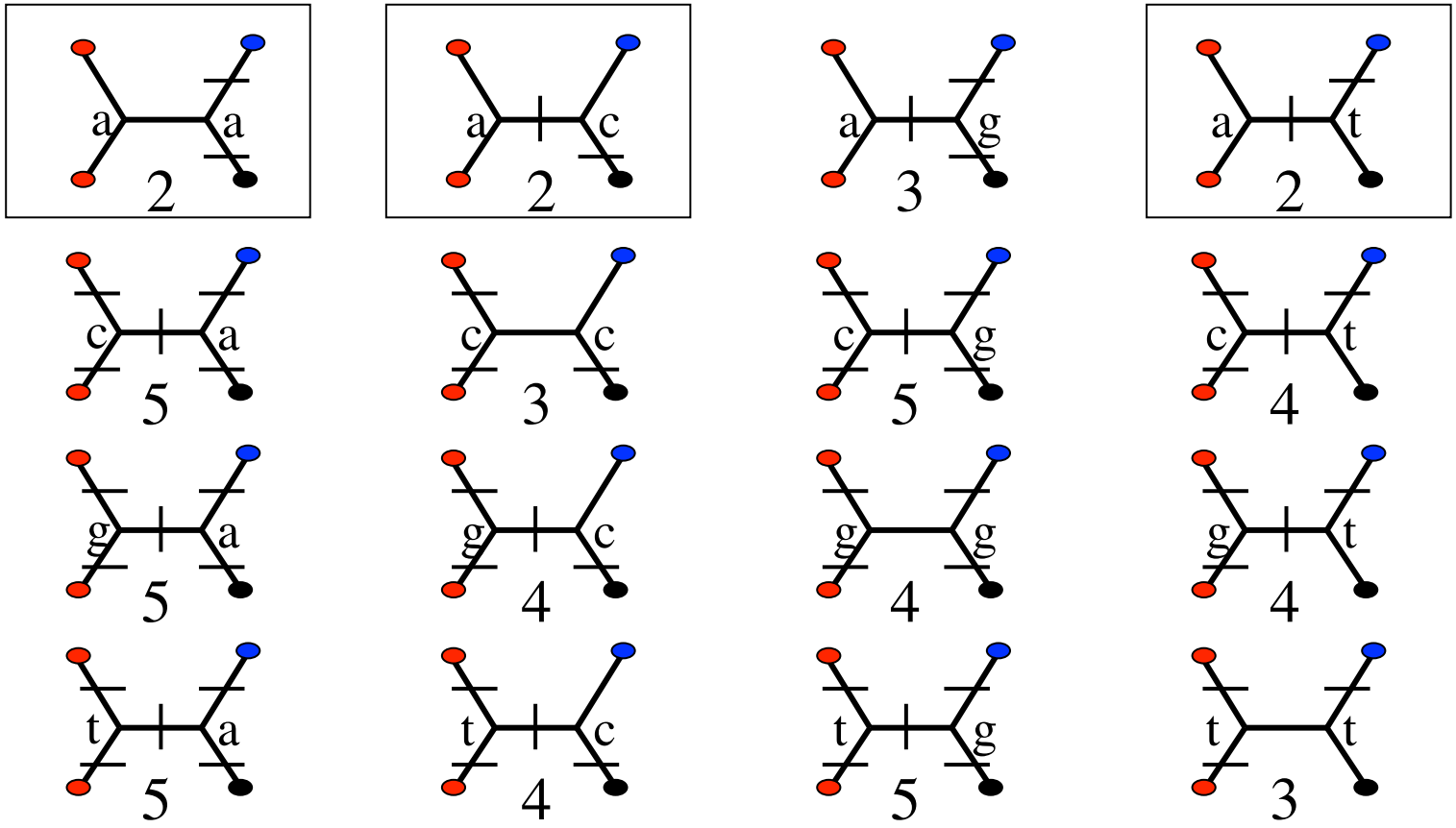
Two Steps

Phylogeny Parsimony

Small problem – brute force

Fitch parsimony (unordered/nonadditive)

$4^{(n-2)}$ internal state sets



Example with PAUP*

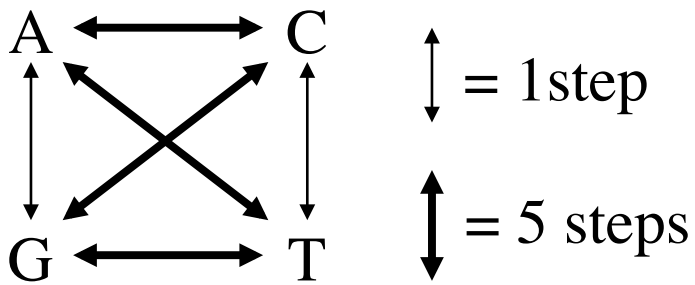
1. Execute primates data set
2. Find the most parsimonious reconstruction

```
execute primate-mtDNA-interleaved.nex;
mprsets 10;
```

Maximum Parsimony Variants

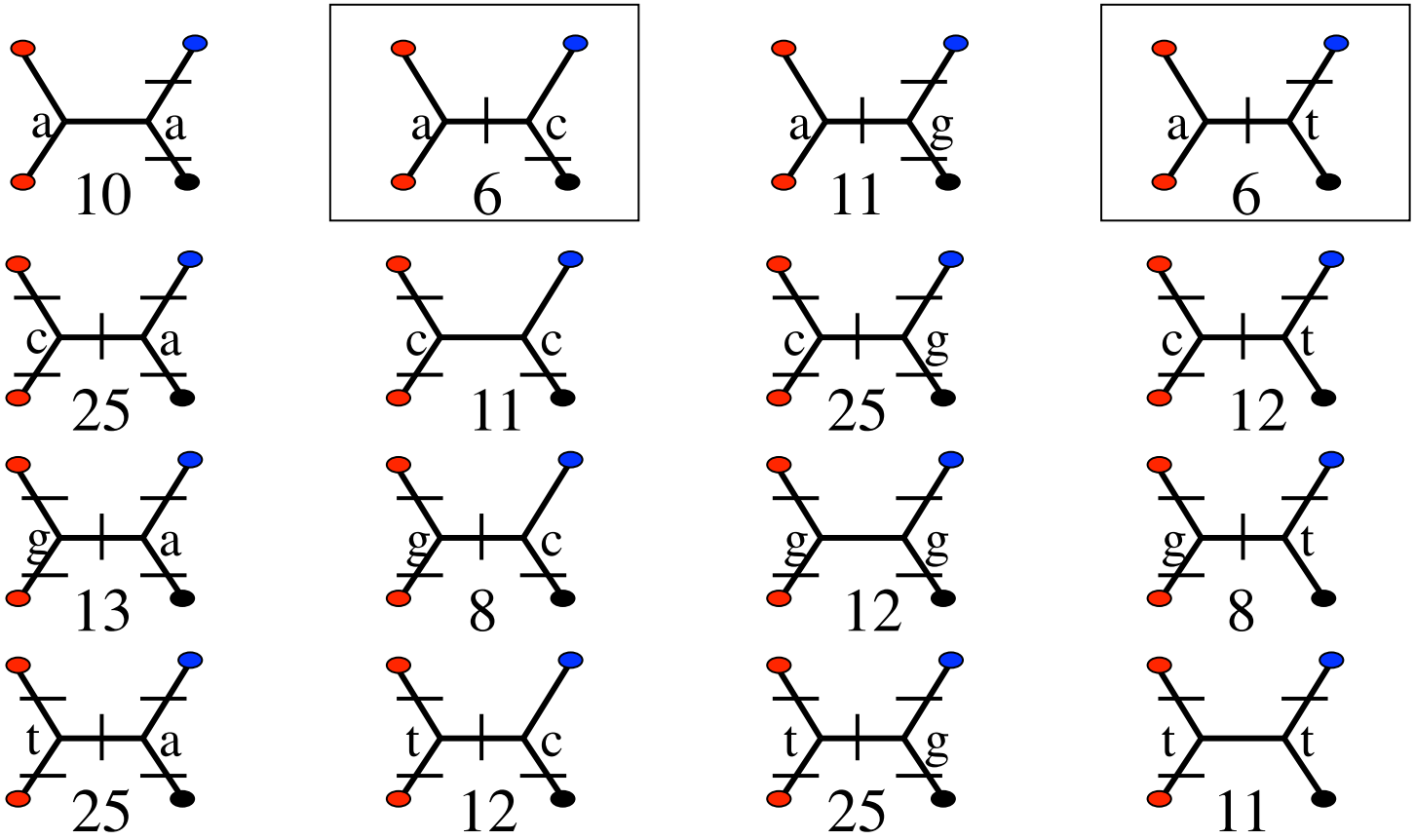
Generalized or Sankoff parsimony

Assign different costs to different changes. For example, Transversions or changes between a purine (A/G) and a pyrimidine (C/T) can be weighted higher than transitions, which are changes between two purines or between two pyrimidines.



Parsimony Variants

Generalized parsimony (Tv=5 and Ti=1)



Example with PAUP*

1. Look at what user types have already been defined.
2. Open the primates data set in the PAUP* editor.
3. Set the transition types to 2_1
4. Look at the possible ancestral character sets for position 10.
5. What is missing in one of the ancestral character sets?

- `showusertypes;`
- `ctype 2_1:all;`
- `mprset 10;`

Previous examples calculated tree lengths under parsimony using "brute force"

- For each character:
  - Consider every possible ancestral state reconstruction
  - Count total cost required for each of these reconstructions
  - Sum over all characters

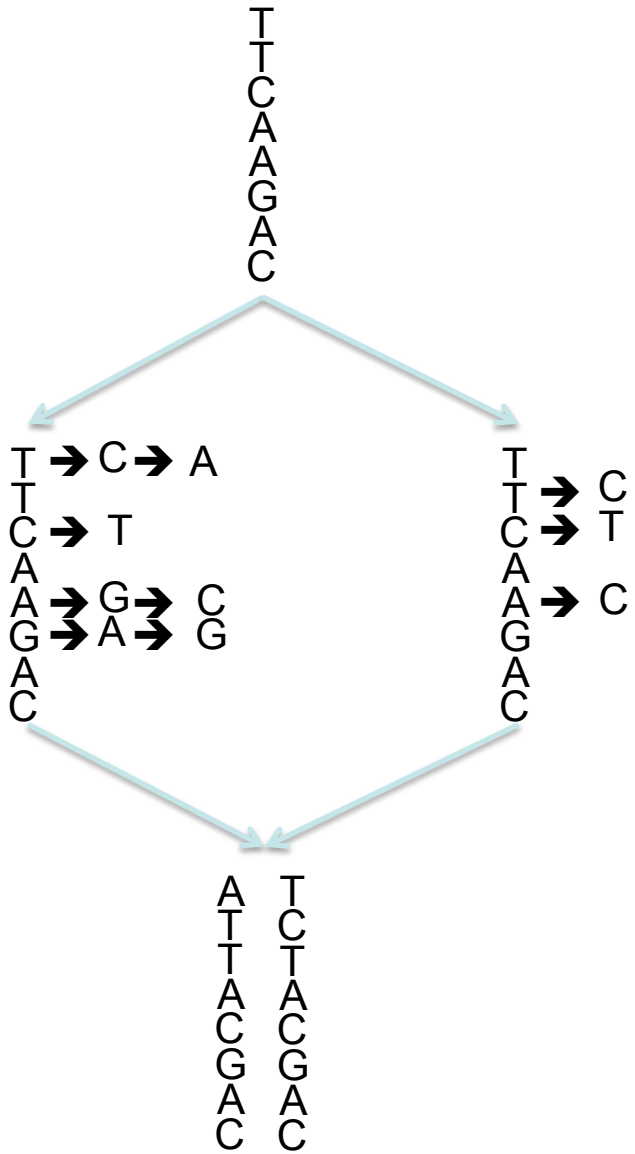
Maximum Likelihood Criterion

as used in phylogenetics

- Formalized for Phylogenetic reconstruction by Joseph Felsenstein in 1981.
- Provides an objective criterion for model comparison.
- Provides a mathematical framework to account for unobserved changes
- Overall goal: Find a tree topology, branch lengths, and associated parameter estimates that maximizes the probability of obtaining the observed data, given a model of evolution.
  - Likelihood(hypothesis)$\propto$Prob(data|hypothesis)
  - Likelihood(tree,model) = k Prob(sequences|tree,model)

Substitution Scenarios

- Multiple substitutions
- Single substitution
- Parallel substitution
- Convergent substitution
- Back substitution



From Yang, 2006: Fig. 1.1

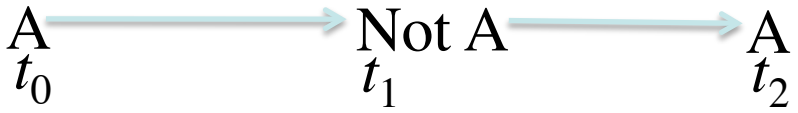
Substitution Models

- For example, what is $p_A(2)$, where $p_A(0) = 1$
- Scenario I: no change

$(1 - 3\alpha)\, p_A(1)$

- Scenario II: changes to non- A and then back to A

$\alpha[1 - p_A(1)]$



- Therefore

$p_A(2) = (1 - 3\alpha)\, p_A(1) + \alpha[1 - p_A(1)]$

- Using recurrence, calculate p for any time interval

$p_A(t+1) = (1 - 3\alpha)\, p_A(t) + \alpha[1 - p_A(t)]$

or

$\Delta p_A(t) = -4\alpha p_A(t) + \alpha$

Substitution Rate Matrix

Jukes and Canter 1969

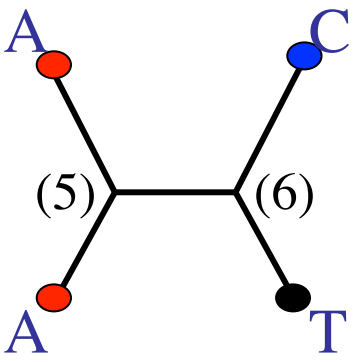
-- Single rate governs all substitutions
-- Alpha represents mean instantaneous substitution rate
-- elements represent rate of change from base *i* to base *j* during inf. dt
-- diagonal are set to minus sum of off diag. so base freqs remain in equilibrium

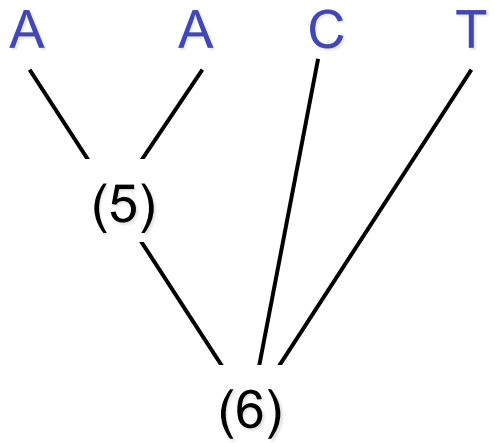
$$
\begin{array}{cc} & \text{To} \\ \text{From} & \begin{array}{c} \\ \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \end{array} \\ \begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} & \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \end{array}
$$

Maximum likelihood inference

in phylogenetics

	1	*j*	*N*
(1)	C…GGACA…	A	…GTTTA…C
(2)	C…AGACA…	A	…CTCTA…C
(3)	C…GGATA…	C	…GTTAA…C
(4)	C…GGATA…	T	…CCTAG…C

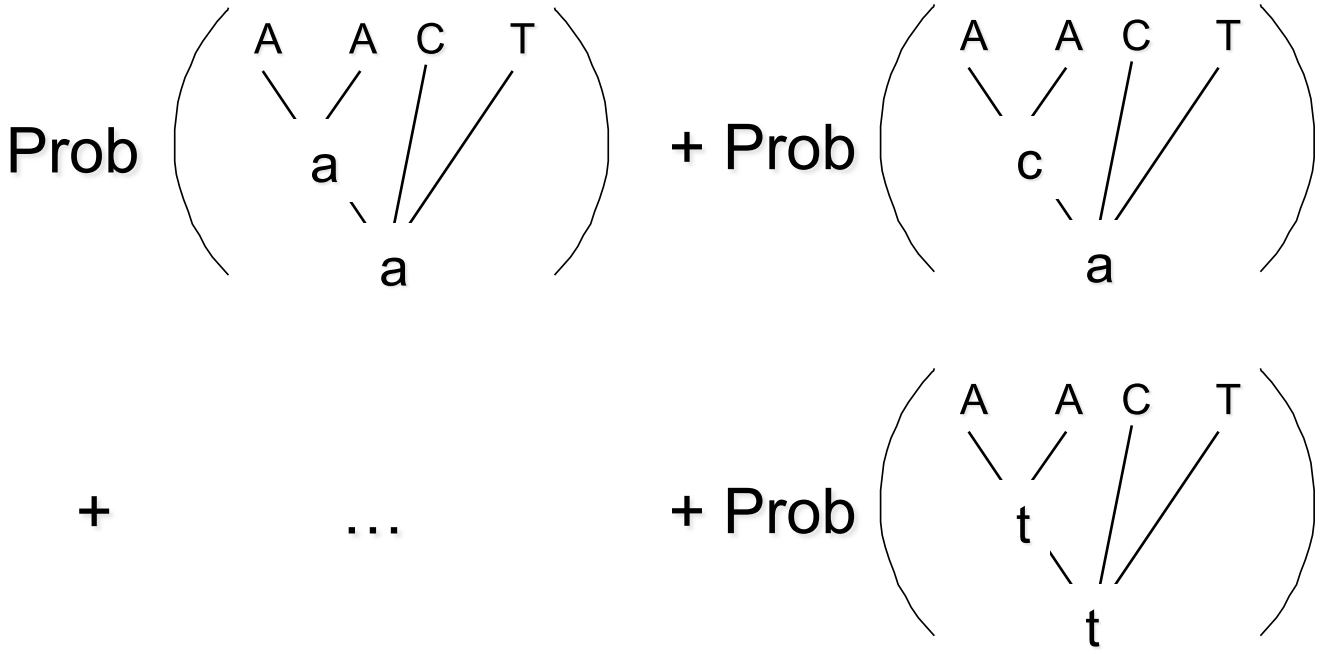




Maximum likelihood inference

in phylogenetics

Log likelihood at site *j* ($\ln L_j$) =

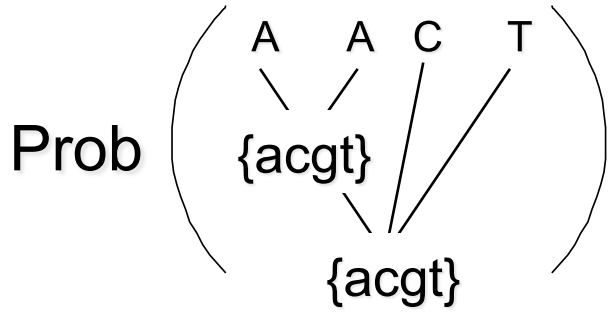


Maximum likelihood inference

in phylogenetics

Single site Score

$$\ln L_j = \mathrm{Pr}_1 + \mathrm{Pr}_2 \ldots \mathrm{Pr}_K = \sum_{i=1}^{K} \mathrm{Pr}_i$$



where K (number of reconstructions) = $4^{(n-2)}$

Total ML Score

$$\ln L = \ln L_1 + \ln L_2 + \ldots + \ln L_N = \sum_{j=1}^{N} \ln L_j$$

Stochastic Modeling

- Continuous-time Markov chains
  - a mathematical model for the random evolution of a memory-less system
  - Used to help solve many multi-dimensional problems
- Properties of Markov model
  - Nucleotide sites evolve independently
  - Four nucleotides are the states of the chain and they are at an equilibrium frequency (stationarity)
  - Substitution rates do not change over time
  - Markov chain has no memory
    - 'given the present, the future does not depend on the past'
  - Assumptions are not always biologically relevant

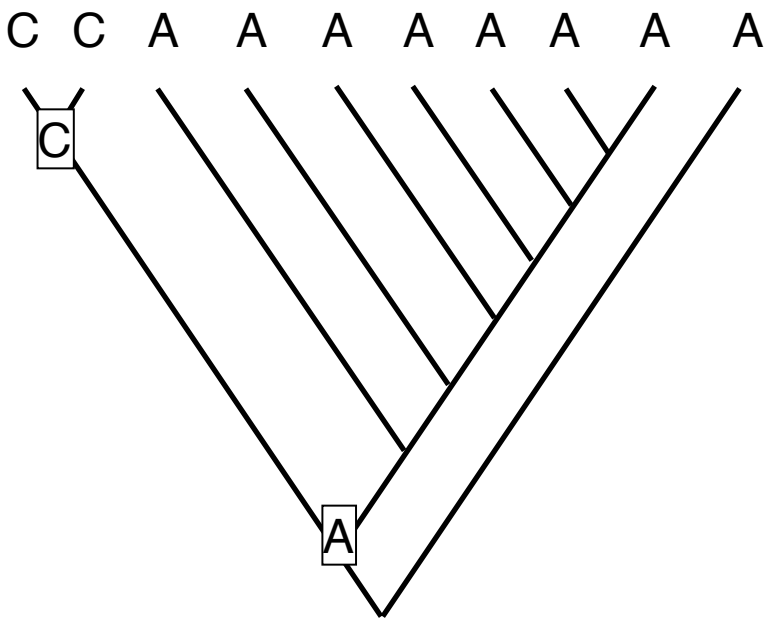
Example with PAUP*

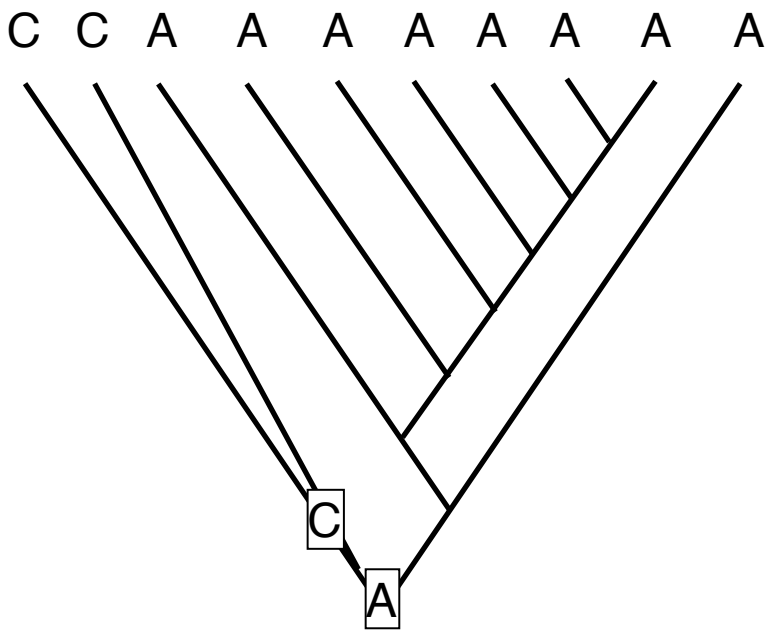
1. Change the optimality criterion to likelihood.
2. What is the best ancestral character reconstruction for character 10?
3. What is the best ancestral character reconstruction for character 1?

```
set criterion=likelihood;
reconstruct 10;
reconstruct 1;
```

Maximum likelihood inference in phylogenetics

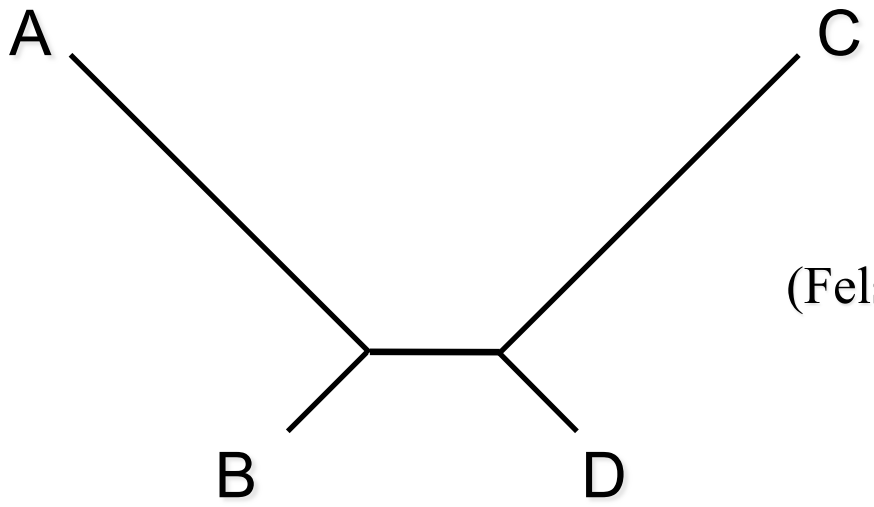
The Relevance of Branch Lengths





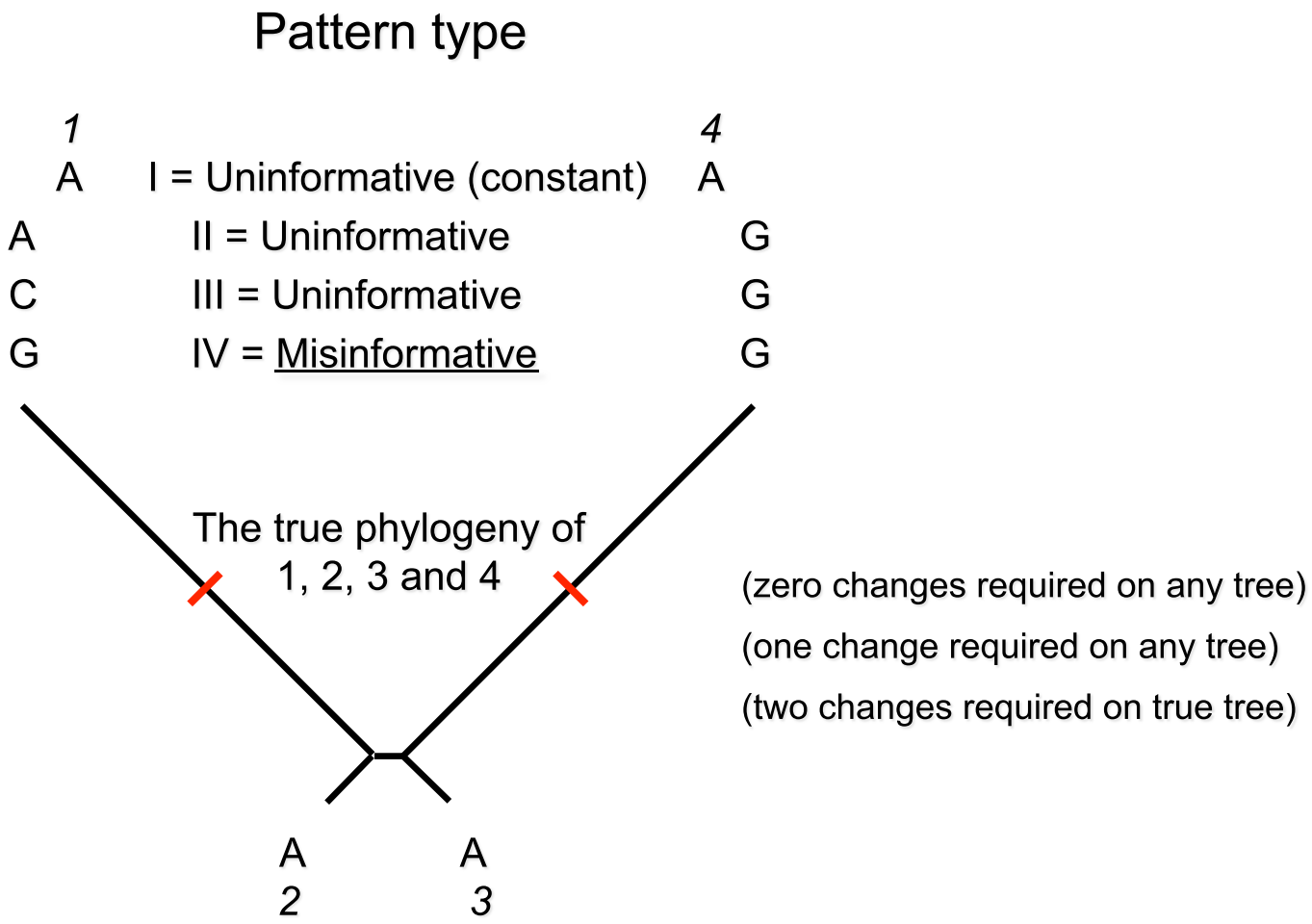
When does maximum likelihood work better than parsimony?

- *When you're in the "Felsenstein Zone"*

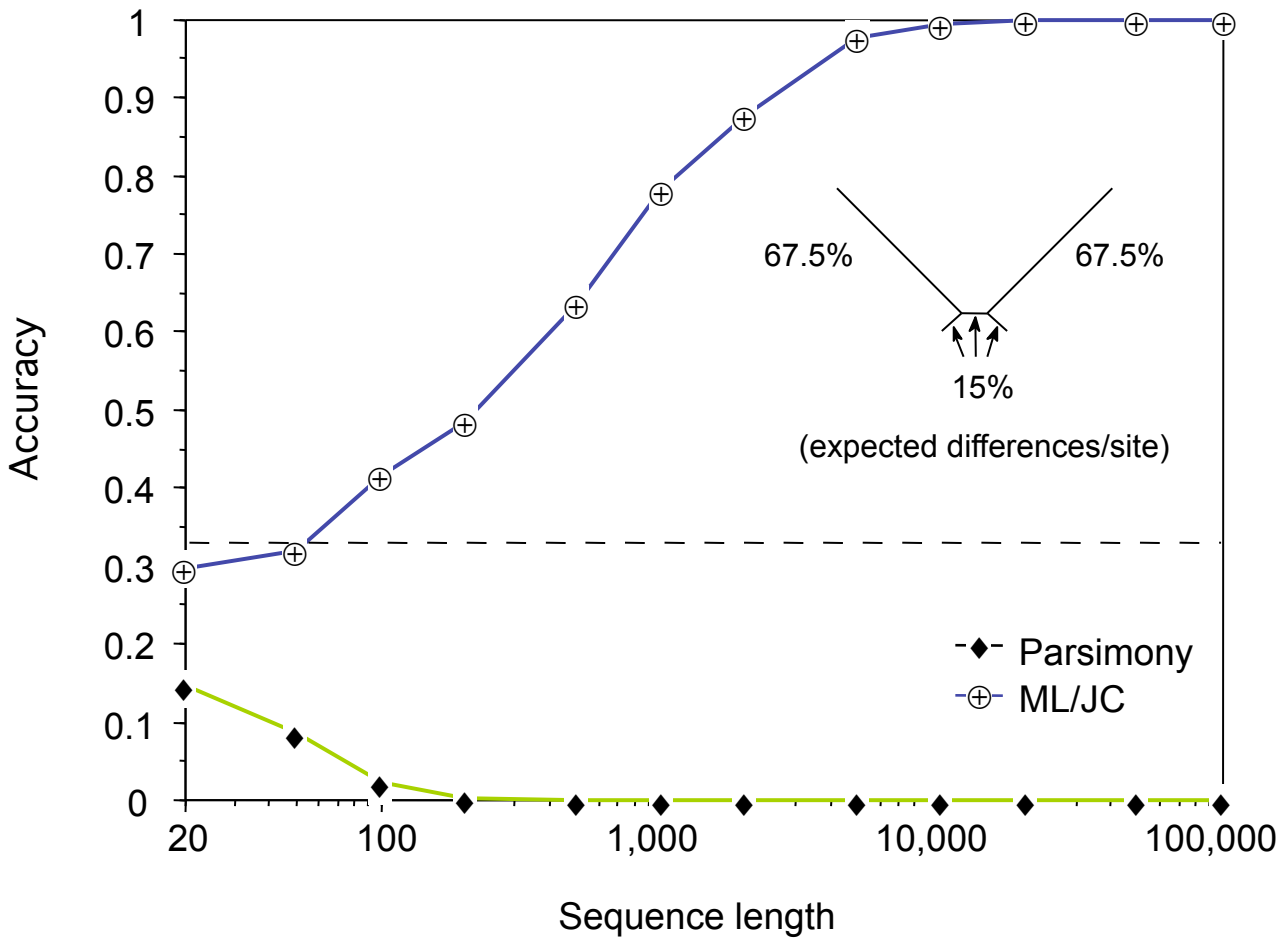


(Felsenstein, 1978)

The long-branch attraction (LBA) problem



Parsimony vs. likelihood in the Felsenstein Zone



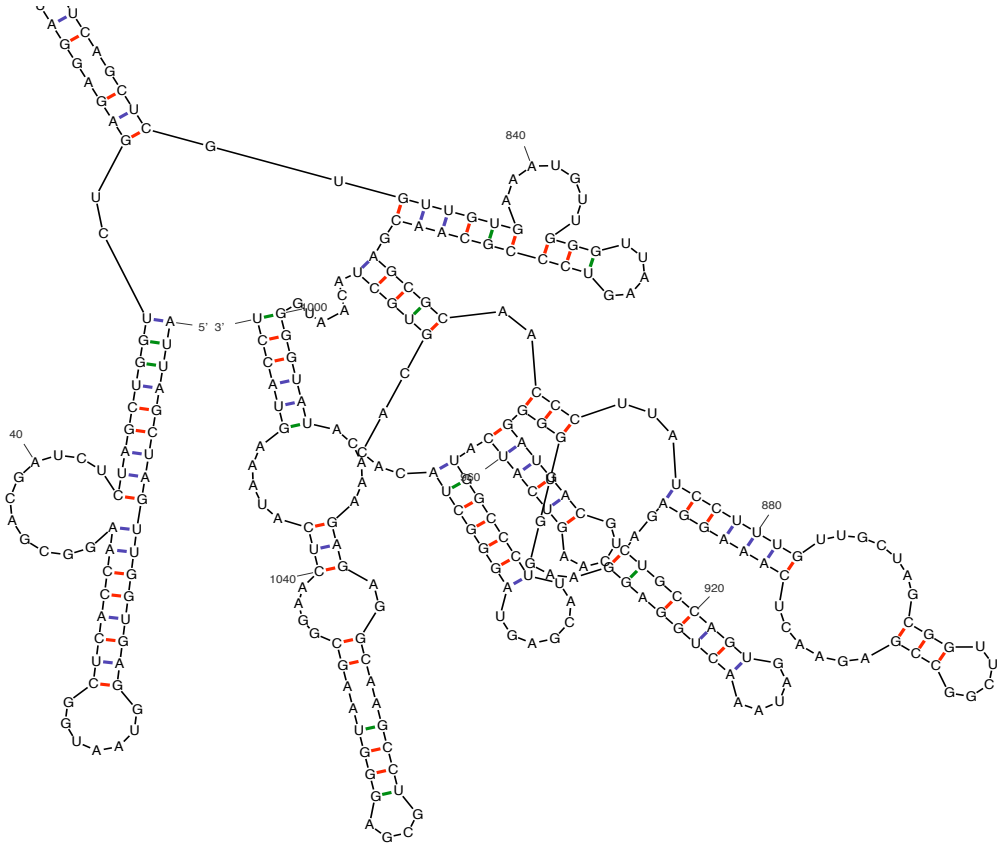
Substitution Models

GTR

General Time Reversible model:
6 substitution types
4 base frequencies

$$
\begin{array}{c}
 \\
\mathrm{A} \\
\mathrm{C} \\
\mathrm{G} \\
\mathrm{T}
\end{array}
\begin{array}{c}
\begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\
\begin{pmatrix}
-r(A\pi_C + B\pi_G + C\pi_T) & \pi_C r_A & \pi_G r_B & \pi_T r_C \\
\pi_A r_A & -r(A\pi_A + D\pi_G + E\pi_T) & \pi_G r_D & \pi_T r_E \\
\pi_A r_B & \pi_C r_D & -r(B\pi_A + D\pi_C + F\pi_T) & \pi_T r_F \\
\pi_A r_C & \pi_C r_E & \pi_G r_F & -r(C\pi_A + E\pi_C + F\pi_G)
\end{pmatrix}
\end{array}
$$

Violations of equal rates



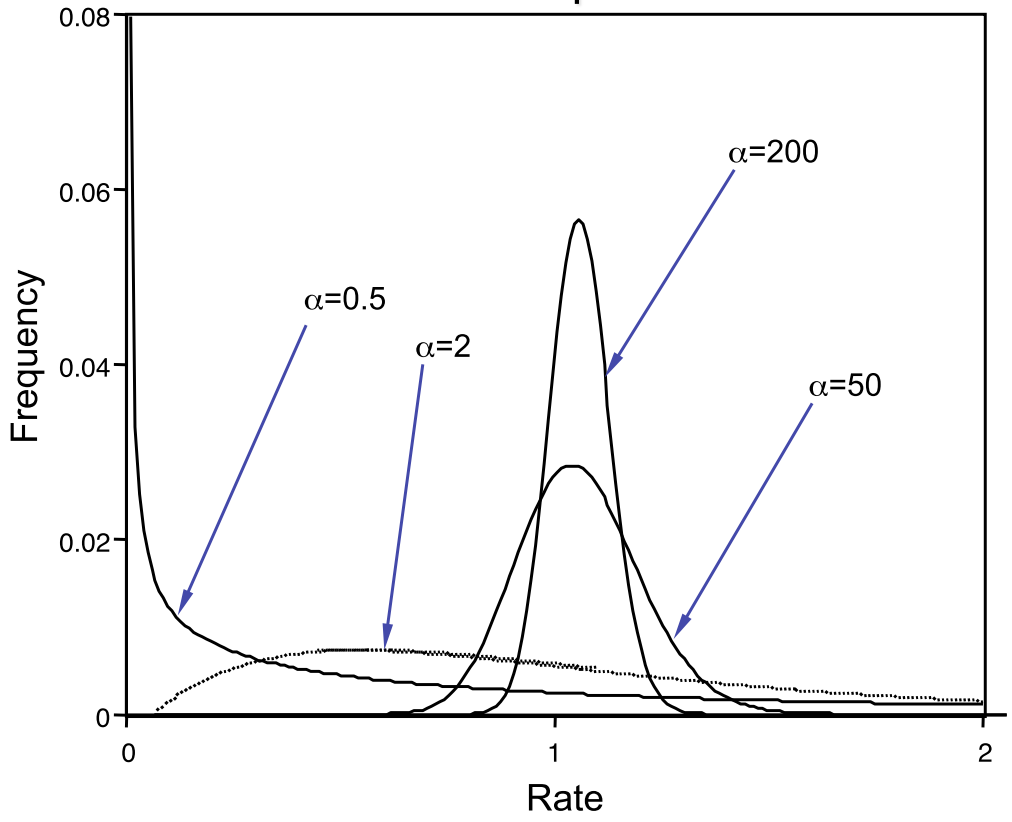
Among site rate heterogeneity


```
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 site don't change do to strong functional or structural constraint (Hasegawa et al., 1985)
- Site specific rates
  - uses a single rate for a sets of characters
- Gamma distributed rates
  - β (scale) and α (shape). Set β to $1/\alpha$ to get a mean rate of 1. (Yang, 1993)

Among site rate heterogeneity

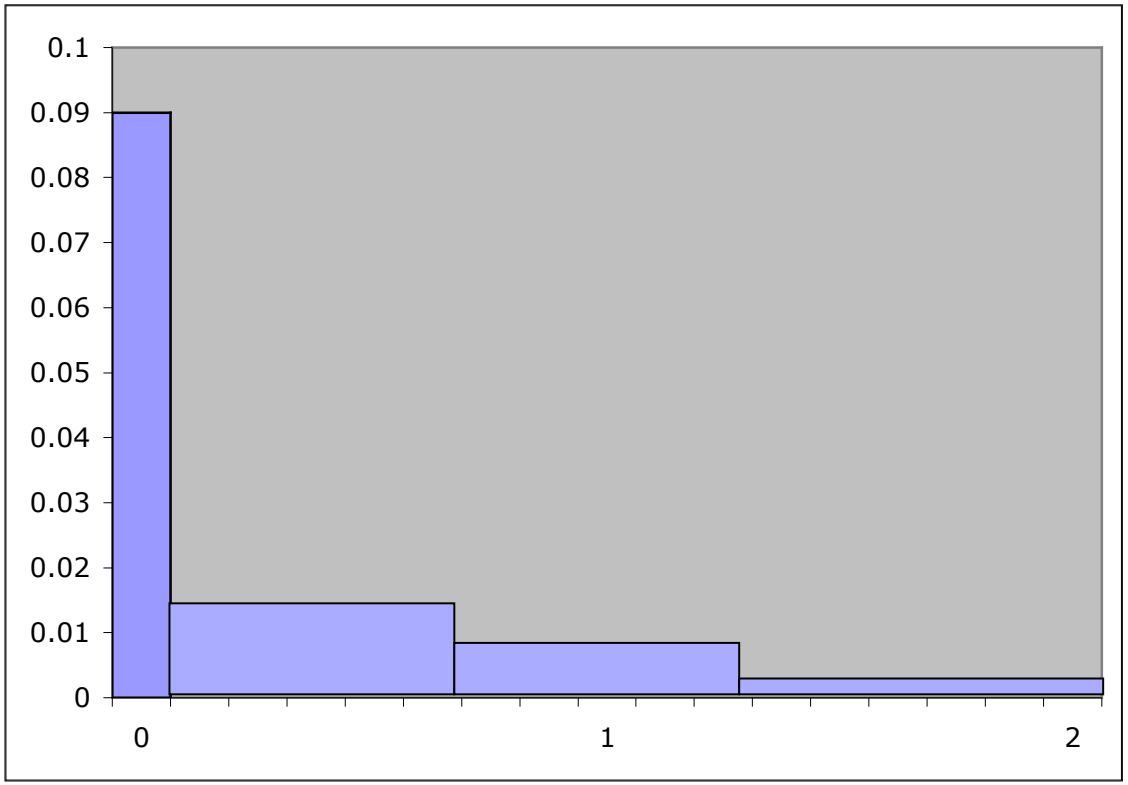
- Model among-site rate variation with a gamma distribution
- Gamma has a scale parameter (β) and shape parameter (α)
- set $\beta = 1/\alpha$ so that mean rate is equal to 1



Among site rate heterogeneity

Modeling among-site rate variation with a gamma distribution...

discrete approximation



Maximum likelihood inference in phylogenetics

Nested Models

GTR+I+Γ 3(bf)+6(sub)+2(rh)	11
GTR+Γ 3(bf)+6(sub)+1(rh)	10
GTR 3(bf)+6(sub)	9
HKY85+I+Γ 3(bf)+2(sub)+2(rh)	7
HKY85+Γ 3(bf)+2(sub)+1(rh)	6
HKY85 3(bf)+2(sub)	5
K2P+I+Γ 2(sub)+2(rh)	4
K2P+Γ 2(sub)+1(rh)	3
K2P 2(sub)	2
JC+I+Γ 1(sub)+2(rh)	3
JC+Γ 1(sub)+1(rh)	2
JC 1(sub)	1

Choosing a model

"Essentially, all models are wrong, but some are useful" *Box, 1987*

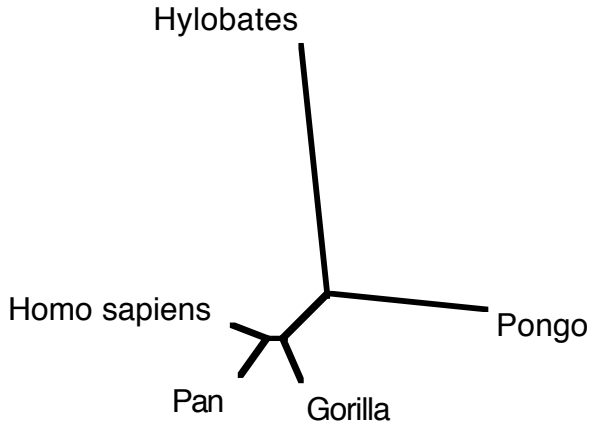
Example with PAUP*

1. Select the General Time Reversible model
2. Estimate the score of the tree currently in memory
3. Add the rate heterogeneity parameter gamma.
4. Estimate the score of the tree currently in memory.
5. What is the difference between the scores?

```
lset nst=6 matrix=estimate;
lscore 1;
lset rates=gamma shape=estimate;
lscore 1;
```

Maximum likelihood inference

Choosing a DNA substitution model



GTR+Γ

$\ln L_1 = -2625.73859$

GTR

$\ln L_0 = -2664.43013$

$38.69 \pm ?$, with one additional parameter

Choosing among Likelihood Methods

Choose the model that maximizes a "goodness of fit" statistic without adding unnecessary parameters.

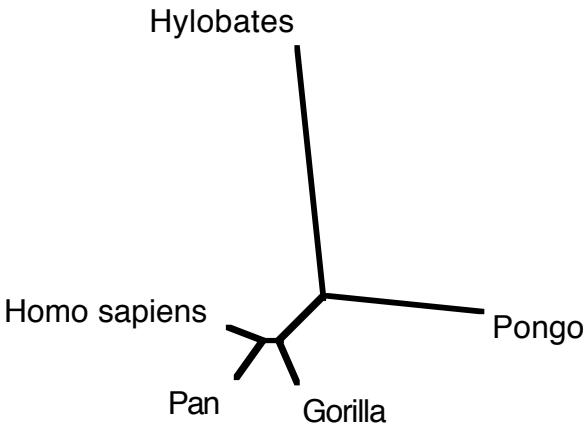
- Likelihood Ratio Test (LRT)
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Monte Carlo Simulations or Parametric Bootstrap

Hierarchical Likelihood Ratio Test

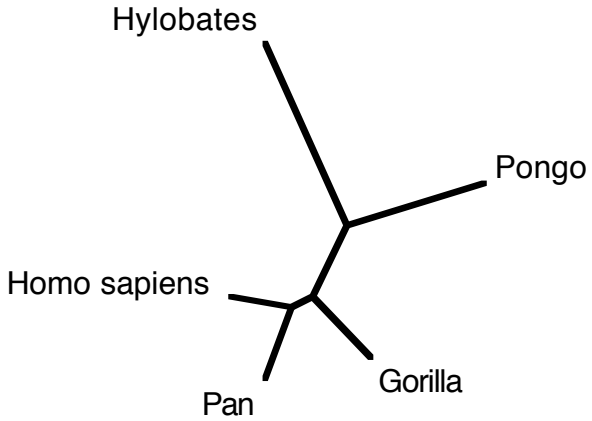
- $\lambda = L_0 / L_1$ choose the simpler model when λ is large
  - L_0 = ML estimate of simpler model (fewer free parameters, lower likelihood -- e.g., without rate heterogeneity).
  - L_1= ML estimate of more complicated model (more free parameters, higher likelihood -- e.g., with rate heterogeneity).
- To obtain a confidence interval we use the fact that $-2\ln\lambda$ is generally χ^2 distributed with k degrees of freedom.
  - where k is the difference between the number of free parameters used to calculate L_0 and L_1 (L_0 has k fewer parameters then L_1).
- Cons
  - Stepwise procedure does not guarantee finding a optimal model.
  - Arbitrary significance level
    - Bigger alpha increase Type I error, reject null when it should have been accepted (Bias toward more complicated model)
    - Smaller alpha increase Type II error, accept null when it should have been rejected (Bias toward simpler model)
  - Problem of multiple tests
  - Starting point dependencies

Maximum likelihood models

for nested models



GTR+Γ



GTR

$-\ln L_1 = 2625.73859$

$-\ln L_0 = 2664.43013$

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

$$\chi^2_{df=1} = 77.38308,\ P < 0.0001$$

Example using JModelTest

1. Compare the two score values using the program JModelTest.
2. Is the difference significant?
3. Are you surprised?

- Open JModelTest;
- Select “Tools”>”LRT calculator”;
- Enter the two likelihood scores;

Akaiki Information Criterion (AIC)

- AIC = $-2\ln L + 2K$
- Where,

  $\ln L$ is the maximum likelihood value of a specific model of nucleotide sequence evolution and tree topology given the data.

  K = the number of parameters free to vary
- Smaller AIC indicates a better model
  - More parameters in model first term decreases and second term increases
- Cons
  - The number of characters in an alignment must be large compared to the number of parameters ($n/K > 40$), otherwise the asymptotic properties of the method are not met.
- Second order or Corrected AIC (AIC_c)
- $AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$

  Where n is the total number of characters in an alignment

  If n is big relative to K then the correction term become negligible.

Bayesian Information Criterion (BIC)

- BIC = $-2\ln L + K\log(n)$
- Where,

  $\ln L$ is the maximum likelihood value of a specific model of nucleotide sequence evolution and tree topology given the data.

  K = the number of parameters free to vary

  n = the total number of characters in an alignment
- Smaller BIC indicates a better model
- If $n > 8$ the BIC selects simpler models than the AIC
- Possible Cons
  - Assumes flat, improper priors
  - Penalizes parameter-rich models more severely

Pros and Cons of each Method

From Posada and Buckley, 2004

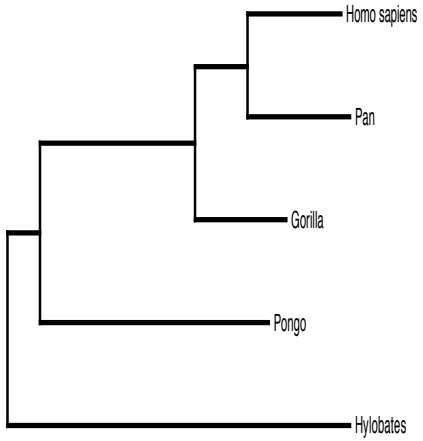
Good properties for model selection methods	hLRT	Bayesian	AIC
Applies easily to non-nested models	no	yes	yes
Simultaneous comparison of multiple models	no	yes	yes
Does not depend on subjective significance level	no	Yes**	yes
Incorporates topological uncertainty	no	yes*	no
Easy to compute	yes	no*	yes
Assesses model selection uncertainty	no	yes	yes
Allows model averaging	no	yes	yes
Possible to specify prior information for models	no	yes	yes
Possible to specify prior information for model params	no	yes*	no
Designed to approximate rather than identify, truth	no	no	yes

*Not the BIC

** In a sense, the interpretation of Bayes factors could be considered as subjective

Parametric Bootstrap:

Monte Carlo simulations



Generate data sets on tree given branch lengths and substitution parameters

Simulated data sets

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

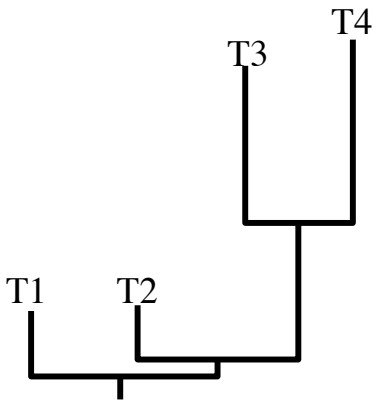
```
Lemur AAGCTTCATAG…TTACATCATCCA
Homo  AAGCTTCACCG…TTACATCCTCAT
Pan   AAGCTTCACCG…TTACATCCTCAT
Goril AAGCTTCACCG…CCCACGGACTTA
Pongo AAGCTTCACCG…GCAACCACCCTC
Hylo  AAGCTTTACAG…TGCAACCGTCCT
Maca  AAGCTTTTCCG…CGCAACCATCCT
```

• • •

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

Parametric bootstrap for non-nested models:

Monte Carlo Simulations



GTR+SS

GTR

Simulate under substitution model

Replicate *i*

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

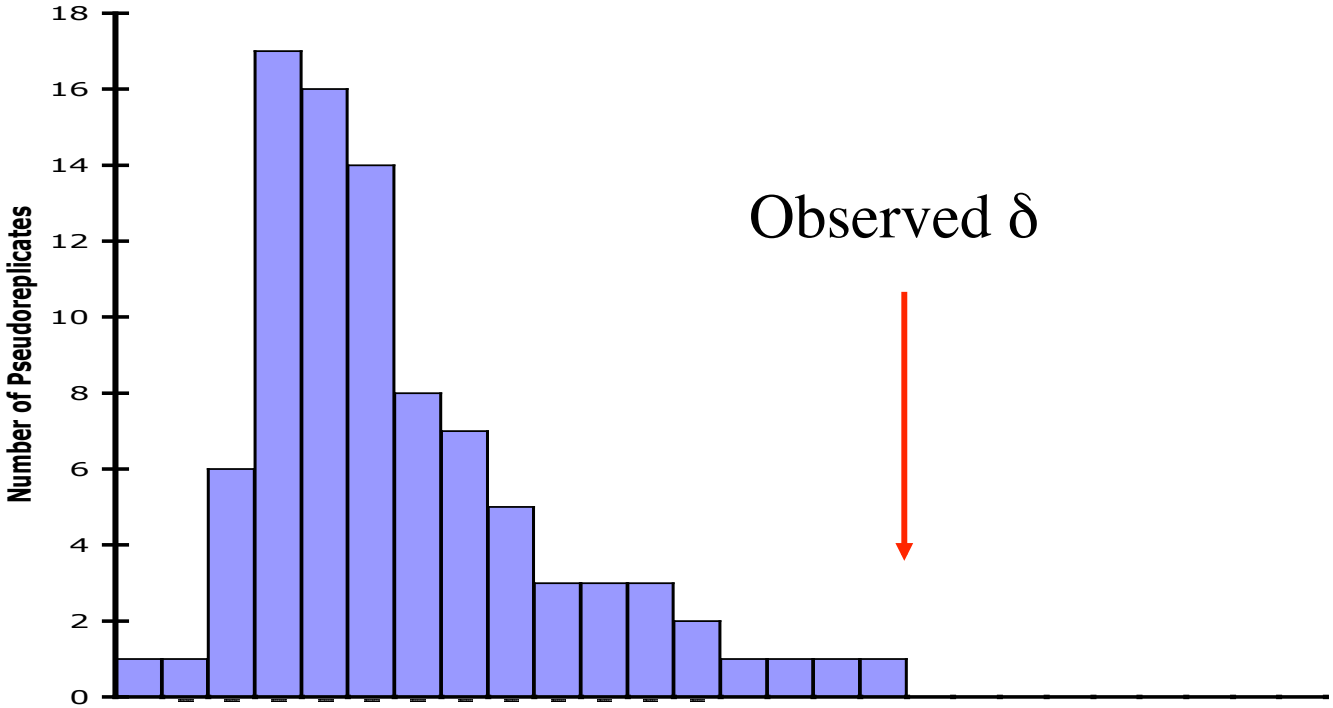
```
Lemur AAGCTTCATAG…TTACATCATCCA
Homo  AAGCTTCACCG…TTACATCCTCAT
Pan   AAGCTTCACCG…TTACATCCTCAT
Goril AAGCTTCACCG…CCCACGGACTTA
Pongo AAGCTTCACCG…GCAACCACCCTC
Hylo  AAGCTTTACAG…TGCAACCGTCCT
Maca  AAGCTTTTCCG…CGCAACCATCCT
```

$\ln L'_1$

$\ln L'_0$

$$\delta'_i = \ln L'_1 - \ln L'_0$$

Parametric bootstrap for non-nested models: Monte Carlo Simulations



$$\delta'_i = \ln L'_1 - \ln L'_0$$

Maximum likelihood inference in phylogenetics

- Uses all the information in the data set
- Requires an explicit model of evolution
- Models are fairly robust to misspecification
- Several goodness-of-fit measures for comparing models

✪ Branch lengths are used in score calculations

Distance as an Optimality Criterion

- Sequence data are reduced to pairwise distances
- Fitch and Margoliash (1967) & Cavalli-Sforza and Edwards (1967)

  Attempts to minimize the difference between the observed pairwise distances and the path length distances between two taxa on a tree topology

$$E = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} w_{ij} \mid d_{ij} - p_{ij} \mid^{a}$$

  d_{ij} is the observed pairwise distance

  p_{ij} is the path length distance between taxa i and j on the tree.

  w_{ij} is a weighting factor that could be used to down-weight distances with high variance.

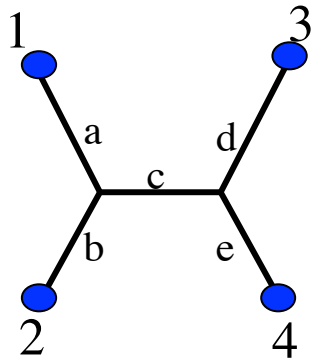
  α is often set to 2, so that E becomes the least-squares fit criterion

- Minimum Evolution (Kidd and Sgaramella-Zonta, 1971)
  - E from FM is used to fit branch lengths
  - the best tree is the one with the smallest sum of branch lengths

Distance as an Optimality Criterion

Least-squares

	1	2	3	4
1	–			
2	d_{12}	–		
3	d_{13}	d_{23}	–	
4	d_{14}	d_{24}	d_{34}	–



$p_{12} = a+b$
$p_{13} = a+c+d$
$p_{14} = a+c+e$
$p_{23} = b+c+d$
$p_{24} = b+c+e$
$p_{34} = d+e$

$p_{ij} = d_{ij}$ for all i and j if the tree topology is correct and distances are additive

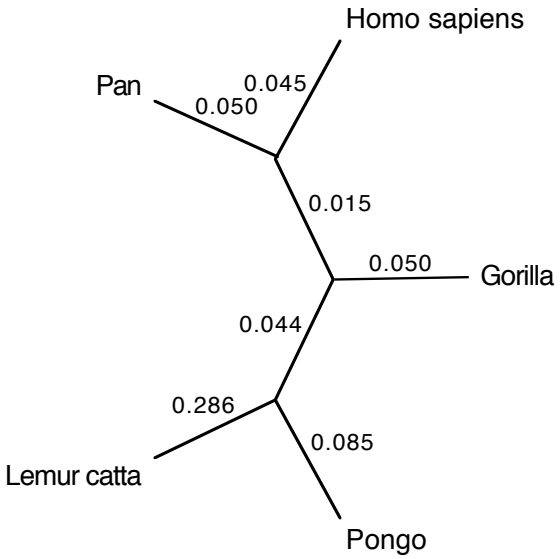
Least-squares branch lengths are those values of a, b, c, d, and e that maximize the fit between p_{ij} and the d_{ij} .

Distance as an Optimality Criterion

Minimum Evolution and Least-squares

ME

LS Brlens
0.28611
0.04436
0.01511
0.04463
0.05044
0.05038
0.08485
0.57588



Least-Squares

d_{ij}	p_{ij}	SS
0.39646	0.39021	0.000039
0.39838	0.39602	0.000006
0.09506	0.09507	0.000000
0.37222	0.38084	0.000074
0.11172	0.11011	0.000003
0.11431	0.11592	0.000003
0.37096	0.37096	0.000000
0.18107	0.18894	0.000062
0.19399	0.19475	0.000001
0.18820	0.17958	0.000074
		0.000261

Example with PAUP*

1. Set the optimality criterion to distance.
2. Calculate the Minimum Evolution (ME) score of the tree
3. What is the distance used for the ME objective function?
4. Change the objective function to least-squares.
5. What is the score of the tree now?

```
set criterion=distance;
dscore 1;
dset objective=lsfit;
dscore 1;
```

Select an Optimality Criterion

- Parsimony
  - discrete characters
  - not explicitly model-based
- Likelihood
  - discrete and continuous characters
  - explicit model of evolution
- Distance (Least-squares and Minimum Evolution)
  - pairwise distances
  - some distances are/are not explicitly model-based

Choosing among methods

Objective Criteria

- Consistency
  - ability of a method to converge on the truth as more data are accumulated
- Efficiency
  - how quickly a method converges on the truth as more data are accumulated
- Robustness
  - measure of sensitivity to violations of a method's assumptions
- Computational speed
  - time required to obtain a solution
- Discriminating ability
  - a measure of proximity to other trees
- Versatility
  - a measure of the kind of information that can be incorporated into an analysis

Choosing among methods

- Analytical results
  - conditions requiring consistency of methods and similarity between methods (no common mechanism)
  - Felsenstein, 1978; Chang, 1996; Rogers, 1997; Tuffley and Steel, 1997
- Simulations
  - "Mutate Data" on the tree according to the model so that number of changes on a branch are proportional to the defined branch length.
  - Huelsenbeck and Hillis, 1993; Gaut and Lewis, 1995; Huelsenbeck, 1995; Bruno and Halpern, 1999; Swofford et al. 2001
- Experimental phylogenies
  - Track the evolution of biological entities.
  - Fitch and Atchley 1985; Atchley and Fitch 1991 [lab mice]; Hillis et al. 1992, 1994 [virus sequences and mutagens]
- Philosophical
  - Only parsimony is consistent with "Popperian falsification" (Popper, 1959)
  - Kluge 1997; de Queiroz and Poe, 2001

Break/Questions/Wakeup

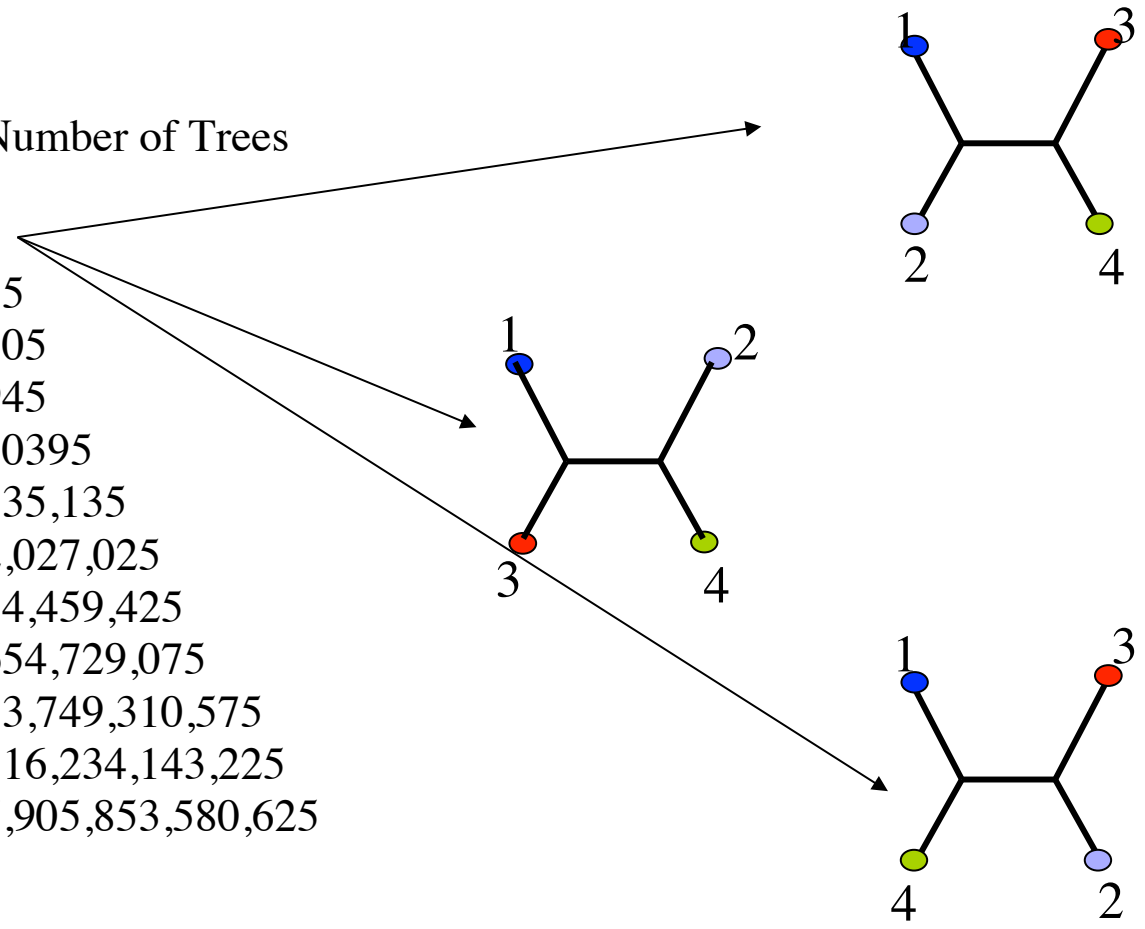
Phylogeny Inference

- What do I want to get out the analyses?
  - Accurate picture of evolutionary relationship among the sequences
  - Comparison among competing hypotheses
  - Estimate of divergence dates
  - Robustness of the data set
- Select an optimality criterion
  - Maximum parsimony
  - Maximum likelihood
  - Distance (Least-squares and Minimum Evolution)
- **Select a search strategy**
- Test the Robustness of the search results
  - Support for individual tree nodes
  - Support for complete tree topologies

Evaluating Trees

Unrooted bifurcating (2N-5)!!

Sequences	Number of Trees
3	1
4	3
5	15
6	105
7	945
8	10395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625



Example with PAUP*

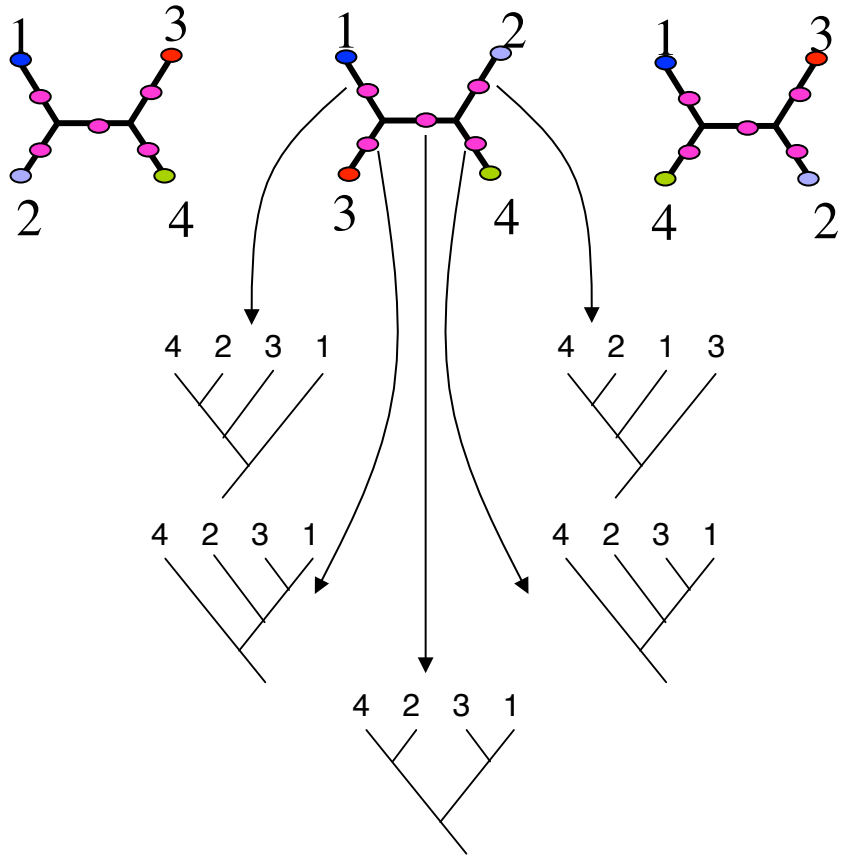
1. Delete all but the first 8 taxa in the primates data set.
2. Change the criterion to parsimony.
3. Calculate the score of all possible trees.
4. How many trees were evaluated?
5. How long did it take?

```
undelete 1-8/only;
set criterion=parsimony;
alltrees;
```

Evaluating Trees

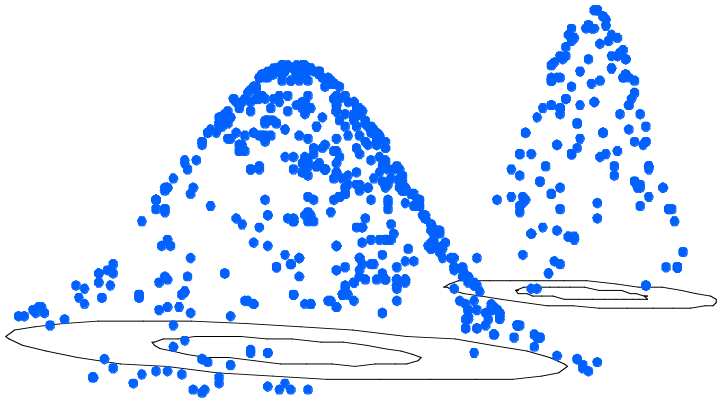
Rooted bifurcating (2N-3)!!

Sequences	Number of Trees
3	3
4	15
5	105
6	945
7	10395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	2,134,580,4667,6875

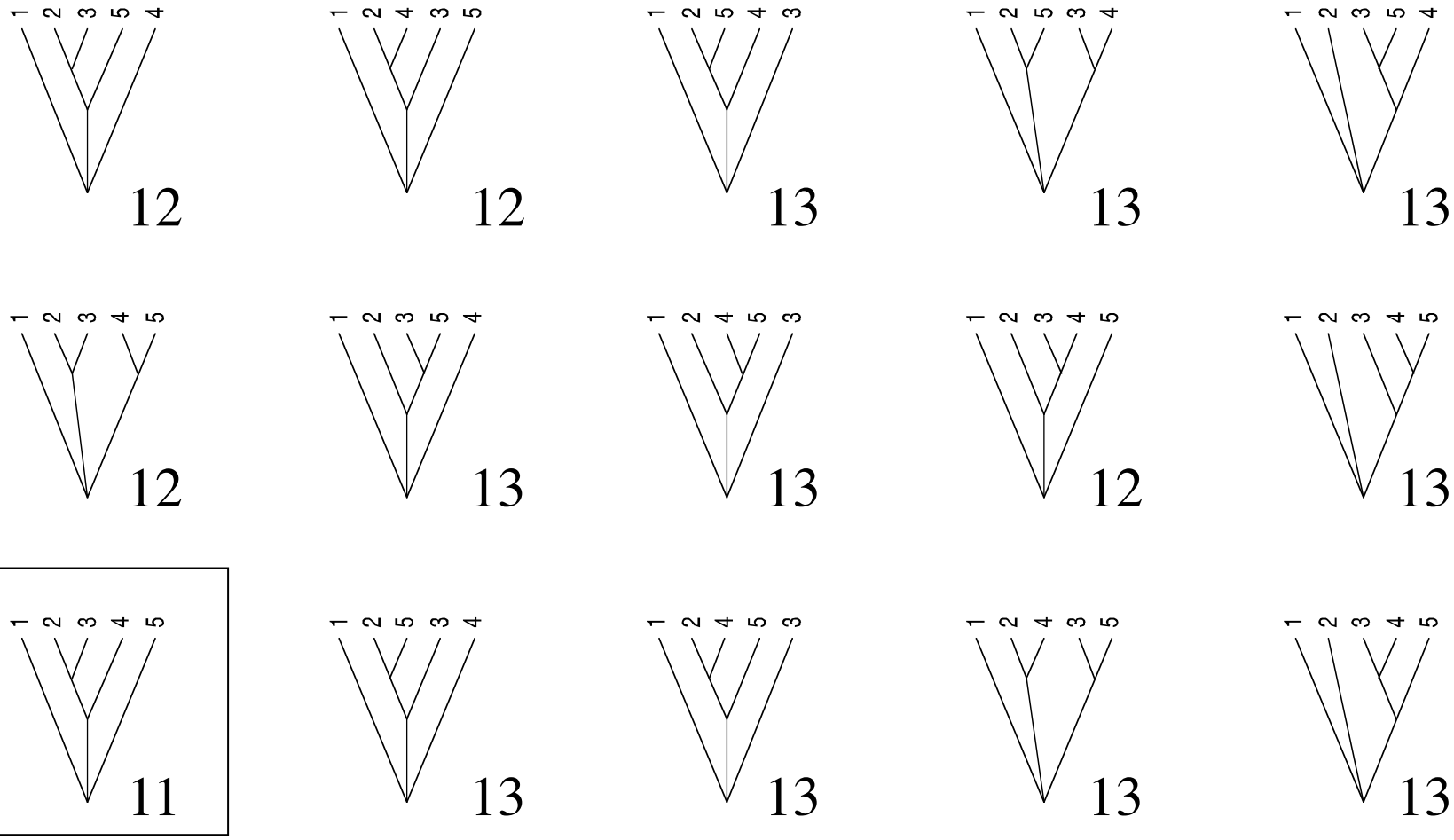


Algorithms for phylogenetic tree reconstruction -- The Large Phylogeny Problem

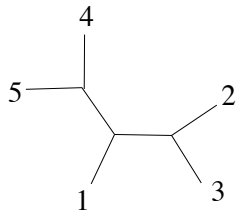
- Exact Methods
  - Exhaustive
  - Branch and Bound
- Heuristic Methods
  - Uphill (greedy) searches
  - Star Decomposition
    - Neighbor Joining
  - Divide and Conquer
    - Short Quartet Method
    - Disk Covering
    - Quartet Puzzling
- Stochastic Methods
  - Simulated Annealing
  - Genetic Algorithms
  - Markov Chain Monte Carlo



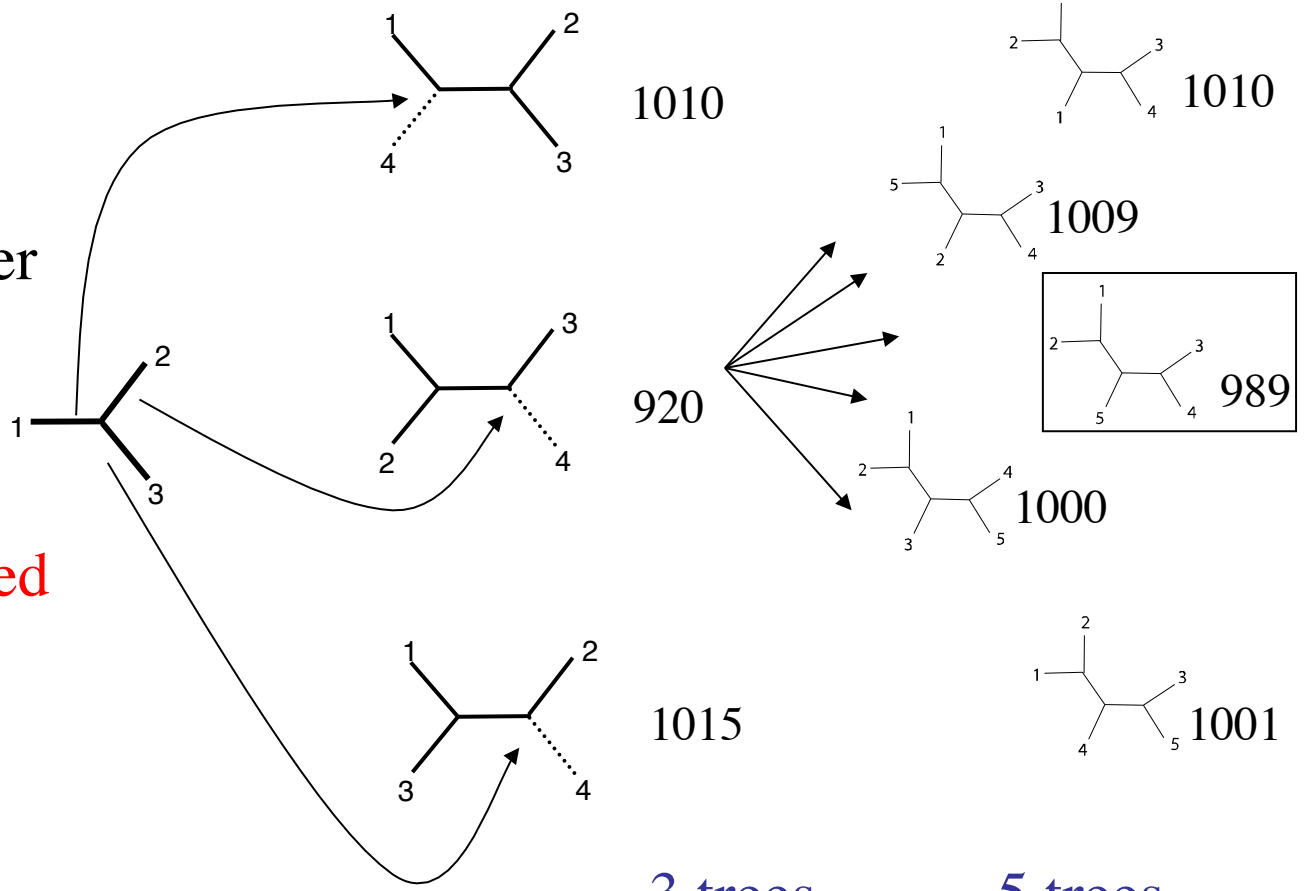
Exhaustive Search



Branch-and-Bound



Get a quick upper bound, say 1001



Only evaluated 8 trees

Example with PAUP*

1. Find the best tree using the "Branch and Bound" algorithm.
2. How long did this take compared to the exhaustive search?

- `bandb;`

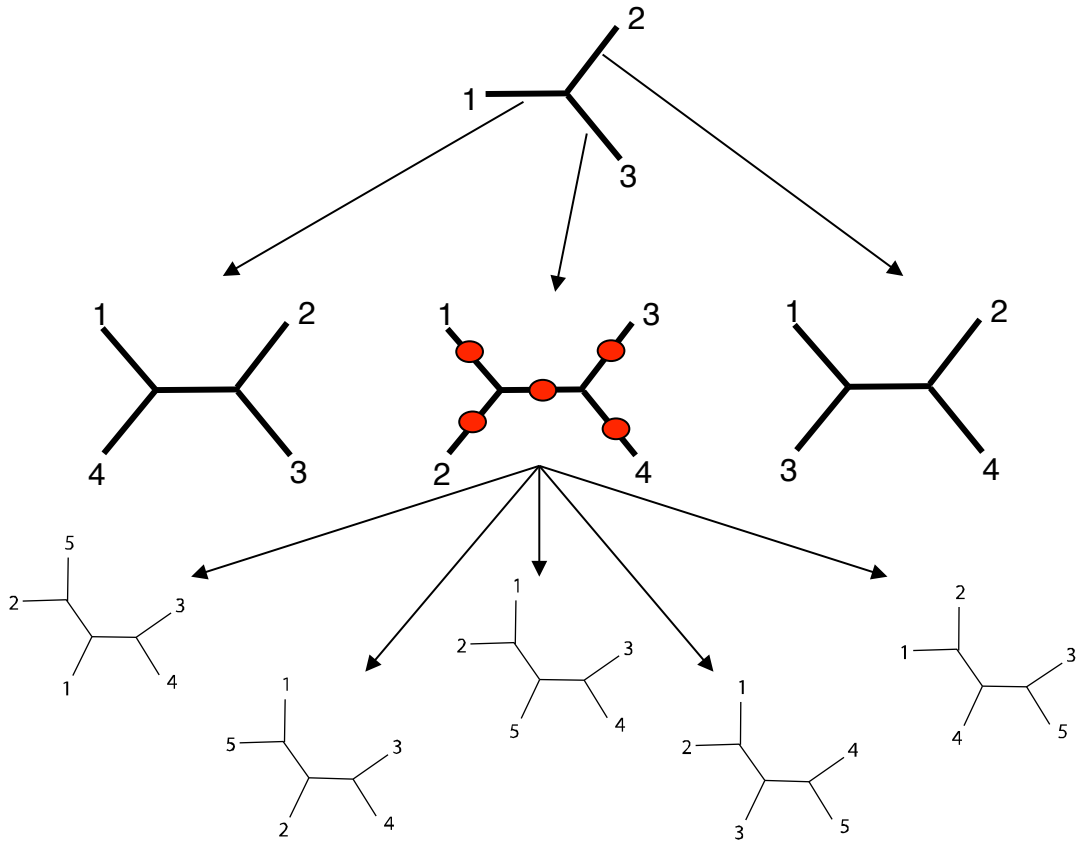
Heuristic Search Algorithms

- Get a starting tree
  - Current tree
  - Neighbor joining
  - Stepwise addition
- Branch swapping
  - Nearest Neighbor Interchange (NNI)
  - Subtree Pruning and Regrafting (SPR)
  - Tree Bisection and Reconnection (TBR)

Getting a Starting Tree

Stepwise Addition Algorithms

E.g., as is, simple, closest, further, random



Getting a Starting Tree

Stepwise Addition Algorithms

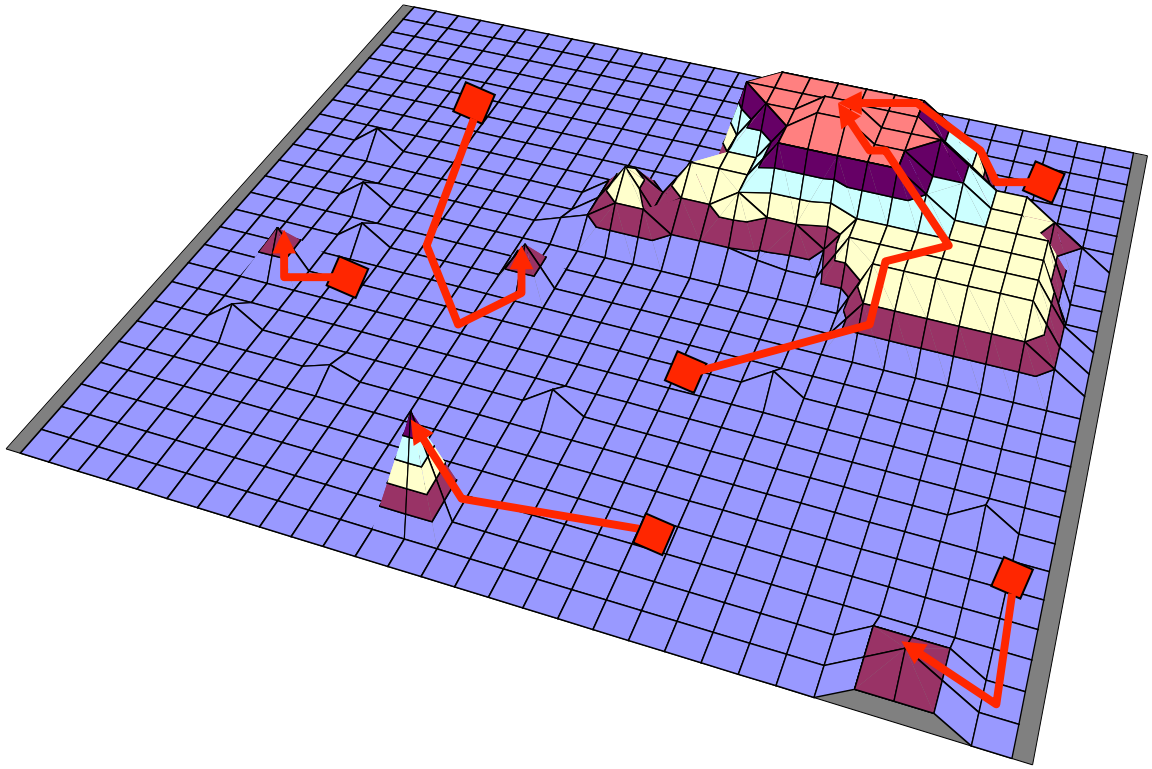
- As Is
  - add in order found in matrix
- Closest
  - add unplaced taxa that requires smallest increase
- Furthest
  - add unplaced taxa that requires largest increase
- Simple
  - Farris's (1970) "simple algorithm" uses a set of pairwise reference distances
- Random
  - random permutation of taxa is used to select the order

Getting a Starting Tree

Stepwise Addition Algorithms

- "Greedy" algorithms
  - only make upward moves
- Prone to getting stuck on local optima

Random Addition Sequence



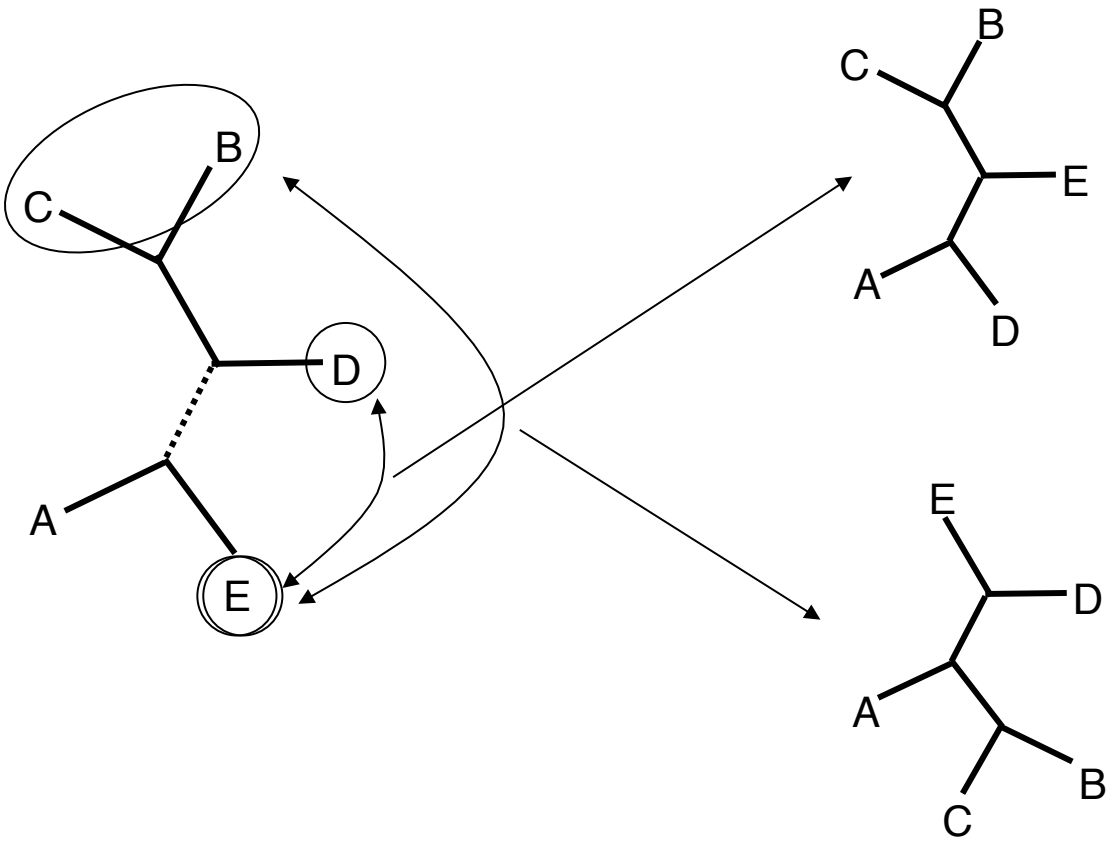
Example with PAUP*

1. Undelete all the taxa in the primates data set.
2. Run a heuristic search
3. What is the default starting tree method?
4. Change the addition sequence method to random.

- `undelete all;`
- `hsearch;`
- `hsearch start=stepwise addseq=random`

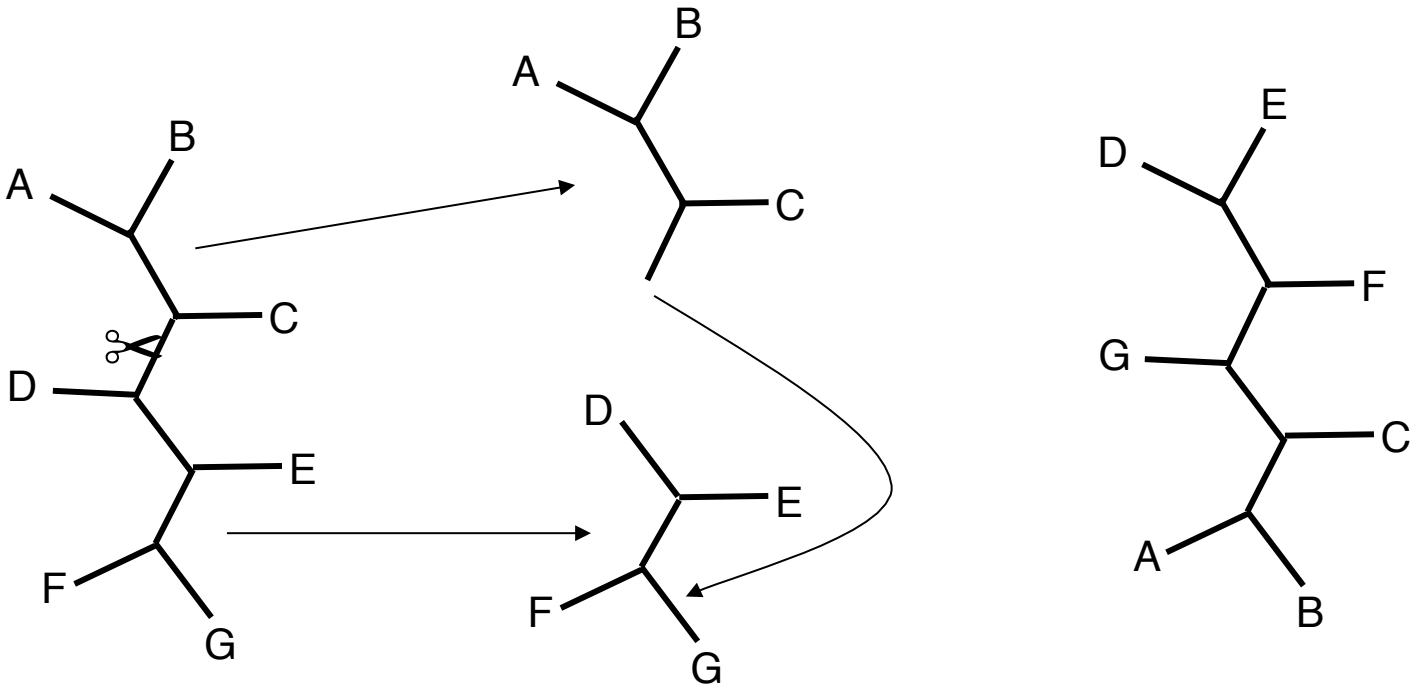
Branch swapping

Nearest Neighbor Interchange (NNI)



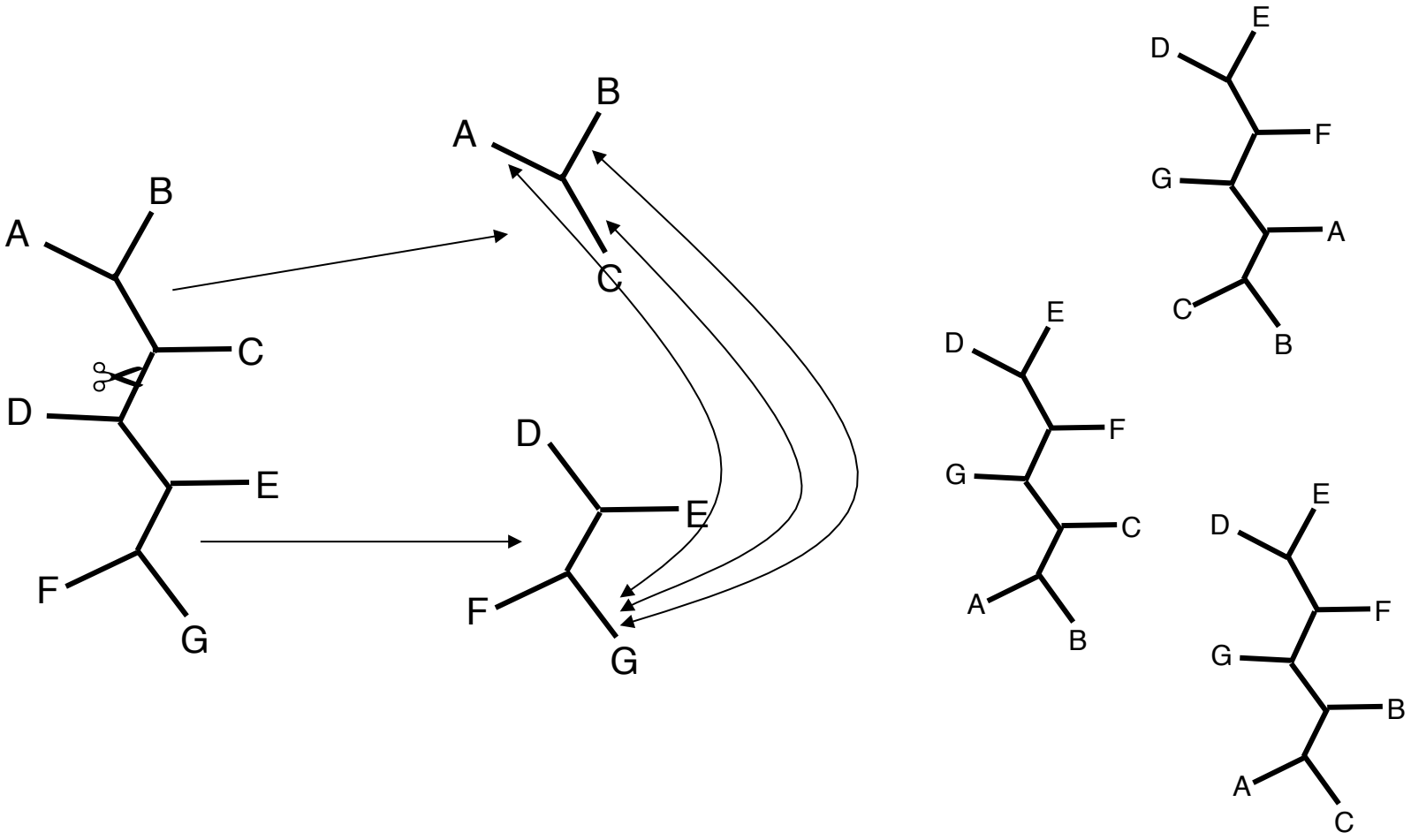
Branch swapping

Subtree Pruning and Regrafting (SPR)



Branch swapping

Tree Bisection and Reconnection (TBR)

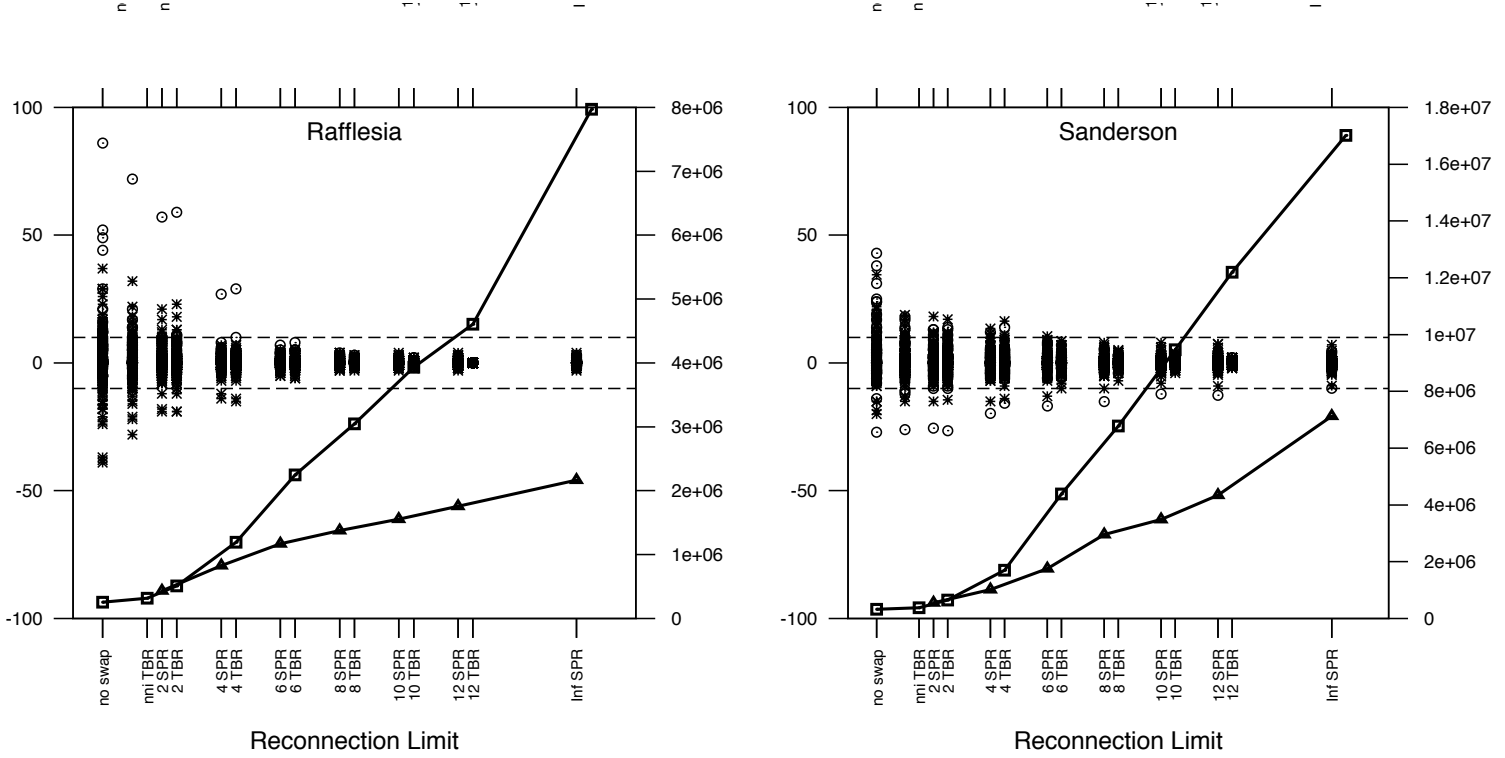


Example with PAUP*

1. Run another heuristic search.
2. What is the default branch swapping algorithm?
3. Change the branch swapping algorithm to nearest neighbor interchange.
4. Does this change reduce the overall runtime of the search?
5. What's the consequence of this change? (hint: how many trees were evaluated?)

- `hsearch;`
- `hsearch swap=nni;`

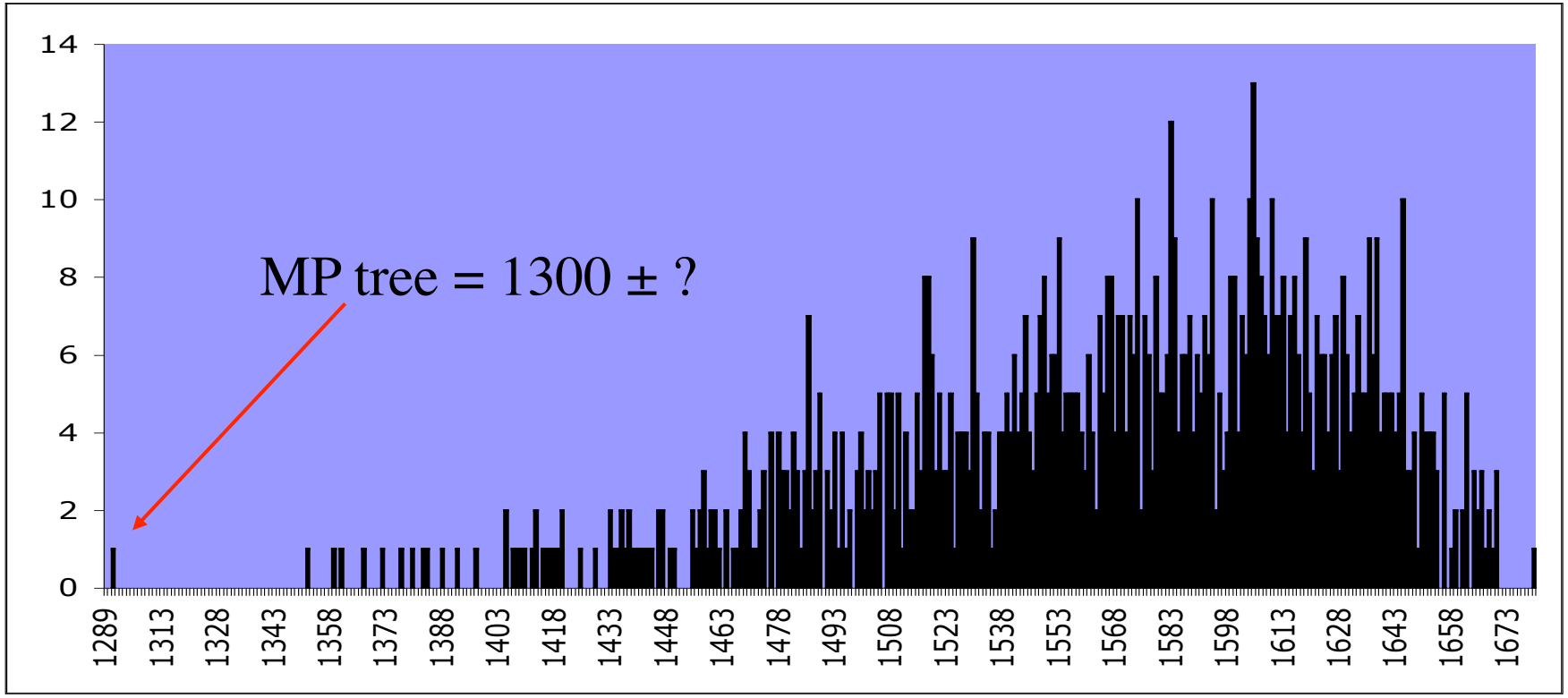
Does Branch Swapping Matter?



Phylogeny Inference

- What do I want to get out the analyses?
  - Accurate picture of evolutionary relationship among the sequences
  - Comparison among competing hypotheses
  - Estimate of divergence dates
  - Robustness of the data set
- Select an optimality criterion
  - Maximum parsimony
  - Maximum likelihood
  - Distance (Least-squares and Minimum Evolution)
- Select a search strategy
- **Test the Robustness of the search results**
  - **Support for individual tree nodes**
  - **Support for complete tree topologies**

Assessing Confidence of the Phylogenetic Tree

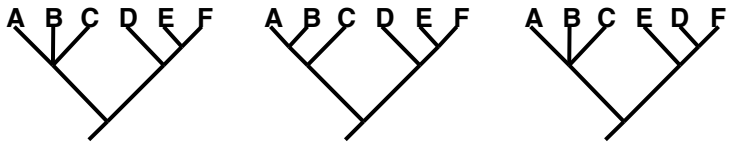


Distribution of tree scores

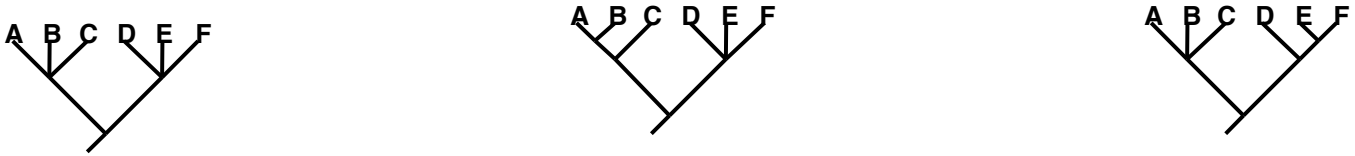
Consensus Trees

- Represent the hierarchical information common to a set of rival trees
  - strict: only those groups appearing on all the rival trees
  - semistrict: groups are retained if they are not contradicted by a rival trees.
  - majority-rule: groups are retained if they are found a pre-specified percentage of rival trees.
  - adams: similar to strict, except that it makes no claim regarding monophyletic groups. Instead, groups are said to be nested.

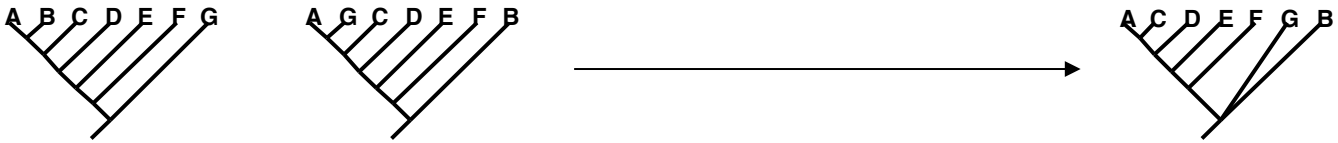
Consensus Trees



1 2 3



strict semistrict 50% majority-rule



1 2 Adams

Example with PAUP*

1. Rest all PAUP* parameters to their default "factory" settings.
2. Run a parsimony heuristic search.
3. Create a consensus tree from the two that your heuristic search found.
4. What is the default consensus method?
5. What information is lost?

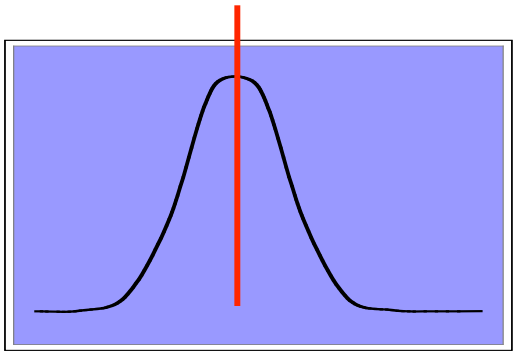
```
reset factory;
hsearch;
contree;
```

Ways of assessing support for a tree topology

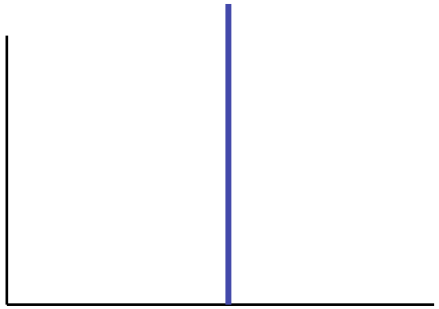
- Bootstrap/Jackknife analyses
- Parametric bootstrap
- KH-test and others
- Bayesian Posterior Probabilities

Bootstrap Technique

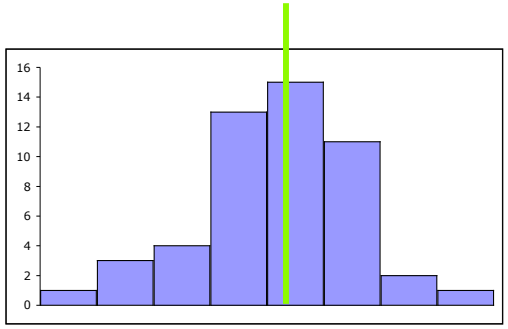
(Efron, 1979)



μ= 5 (true mean)

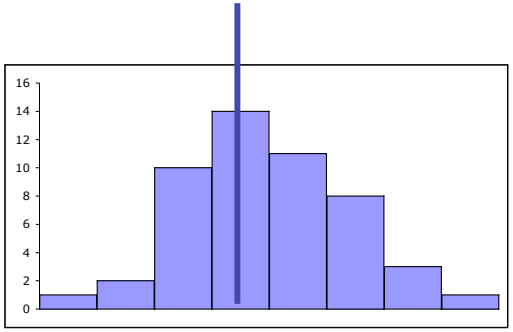


...



$\hat{u} = 5.3$

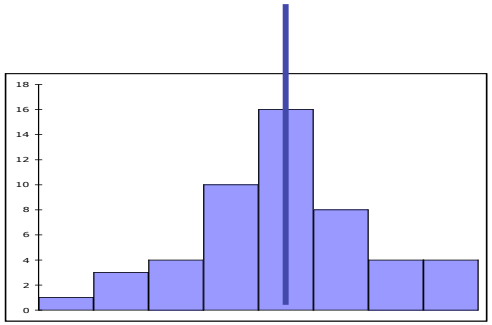
Sample mean
one replicate



$\hat{u} = 4.8$

pseudo replicate 1

...

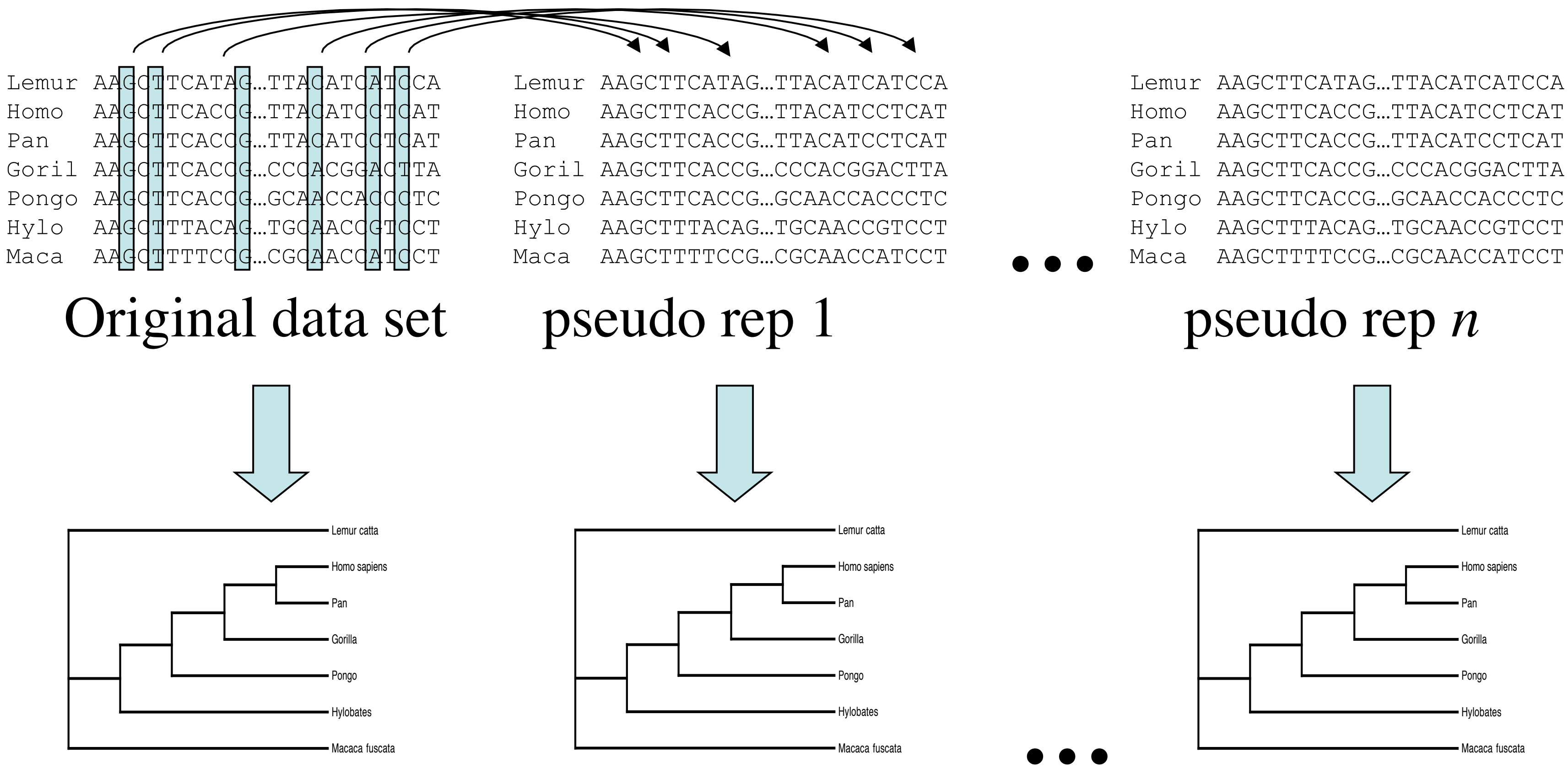


$\hat{u} = 5.1$

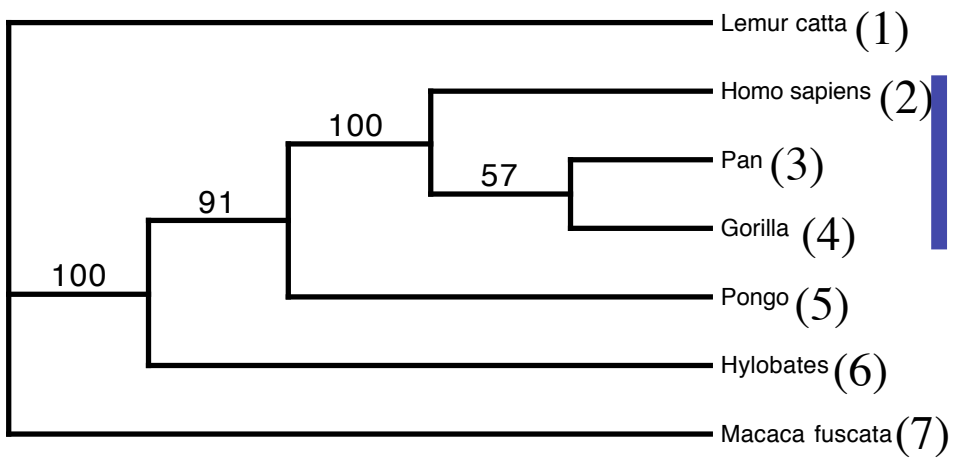
pseudo replicate n

Bootstrapping Phylogenetic Data

(Felsenstein, 1985)



50% Majority-rule Consensus tree




```
1234567          Freq
---------------------
.***...        100.00
.*****.        100.00
.****..         91.17
..**...         57.33
.**....         42.00
.***.*.          7.83
```

Jackknifing Phylogenetic Data

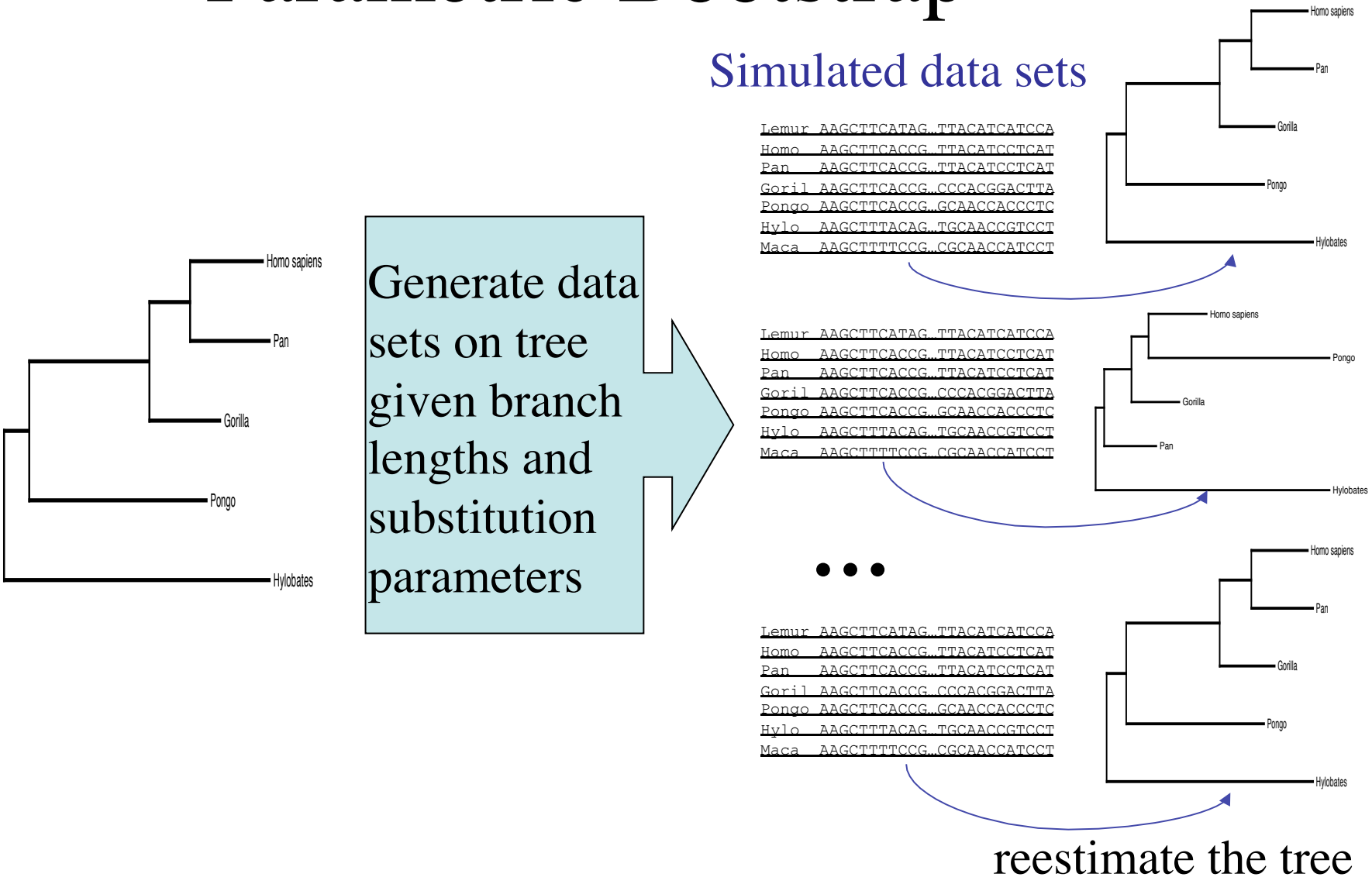
- Also used to assess support for nodes on a given tree
- Data are sampled without replacement
- Replicates represent some fraction of the total data set.
- Jackknife tree is also displayed as a majority-rule consensus tree, where support for a node is given as the percent of the jackknife replicates which contain the split.

Example with PAUP*

1. Switch criterion to maximum likelihood.
2. Run a bootstrap analysis using the default options.
3. How long did this take?
4. Run a bootstrap analysis using the "Subtree Pruning and Regrafting" algorithm and set the reconstruction limit to 8.
5. How long did this take relative to the default heuristic method?

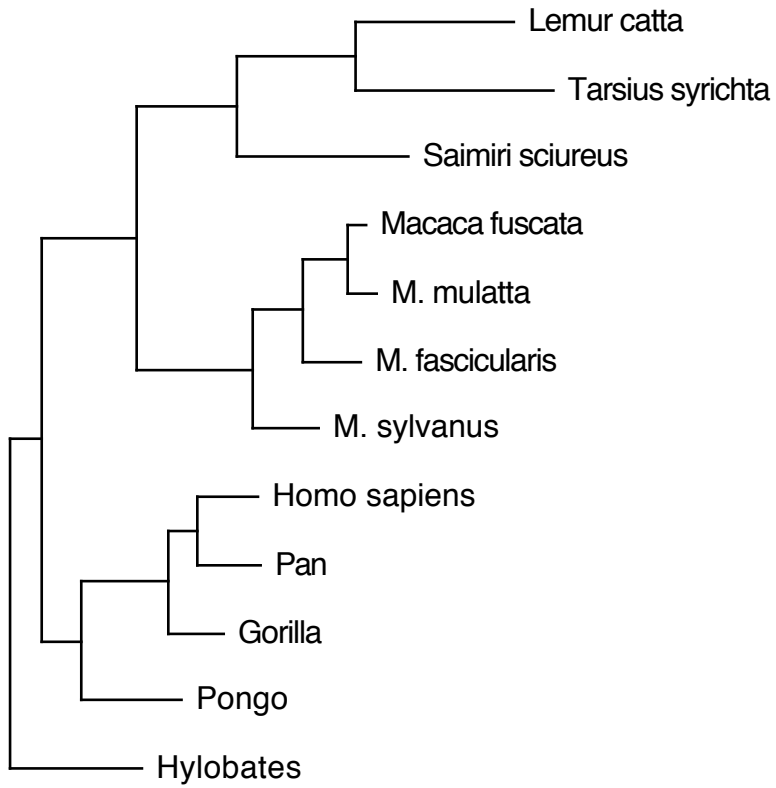
```
set criterion=likelihood;
bootstrap;
bootstrap /swap=spr reconlimit=8;
```

Parametric Bootstrap

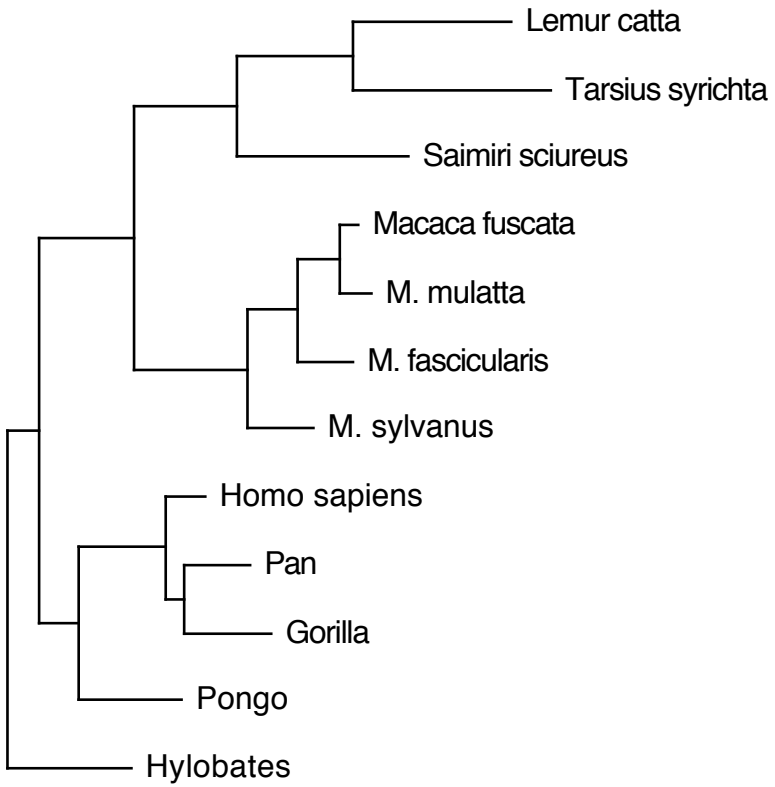


Kishino-Hasegawa Test (KH-test)

(Kishino and Hasegawa, 1989)



-lnL = 5728.06210



-lnL = 5735.81631

7.75420 ± ?

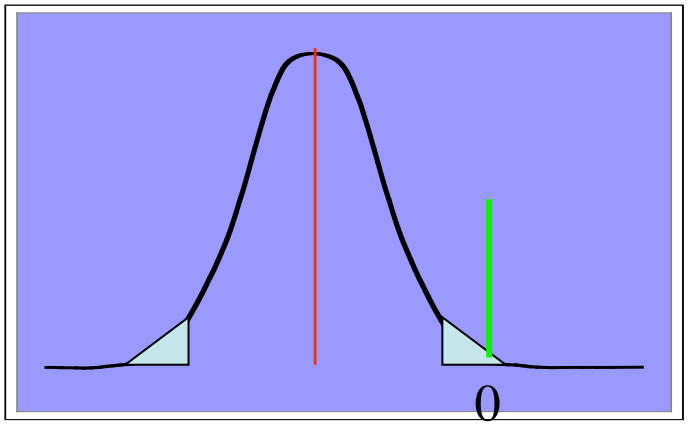
Kishino-Hasegawa Test (KH-test)

Null Hypothesis

If we select the tree topologies a priori:

H_0: $E[\delta] = 0$

H_1: $E[\delta] \neq 0$



The test statistic is the score difference between the competing hypotheses.

What is the distribution of δ under the null?

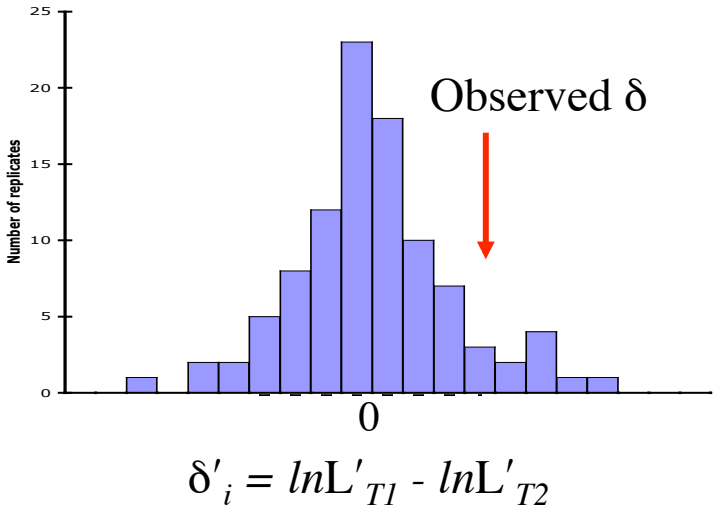
Kishino-Hasegawa Test (KH-test)

Bootstrap Null distribution
(Hasegawa and Kishino, 1989)

Nonparametric bootstrap can be used to generate the Null *F*.
Sample from the empirical data set with replacement

$$\delta'_i = lnL'_{T1} - lnL'_{T2}$$, where *i* is a bootstrap replicate

$$\delta'_1 = lnL'_{T1} - lnL'_{T2}$$
$$\delta'_2 = lnL'_{T1} - lnL'_{T2}$$
$$\delta'_3 = lnL'_{T1} - lnL'_{T2}$$
…
$$\delta'_n = lnL'_{T1} - lnL'_{T2}$$



where *n* is the number of bootstrap replicates

Kishino-Hasegawa Test (KH-test)
(Time saving methods)

- RELL - Resample **E**stimated **L**og-**L**ikelood (Kishino et al., 1990)
  - Replicate *ln*L scores for each tree are obtained by resampling the sitewise log-likelihood scores from the original analysis
  - large data sets required
- Estimate the variance of δ by estimating the variance of the sitewise log-likelihoods (Kishino and Hasegawa, 1989)
  - No resampling is required

$$\nu^2 = \sum_j \frac{(\sigma_j - \bar{\sigma}_j)^2}{n-1}$$

Kishino-Hasegawa Test (KH-test)

(Kishino and Hasegawa, 1989)

```
Kishino-Hasegawa test:
  KH test using RELL bootstrap, two-tailed test
  Number of bootstrap replicates = 1000
KH-test

    Tree          -ln L    Diff -ln L        P
-----------------------------------------------
       1    5728.06210         (best)
       2    5735.81631        7.75420   0.177
```

KH-test Assumptions

- Trees must be selected a priori
- Sites are independently and identically distributed
- Large number of sites are sampled
- Alternative to KH-test relaxes constraint that trees are selected a priori
  - SH-test (Shimodaira and Hasegawa, 1999)

Shimodaira-Hasegawa (SH-test)

(Shimodaira and Hasegawa, 1999)

The test statistic is the score difference between the Maximum Likelihood tree and every other tree compared:

i.e., $\delta_T = lnL_{ML} - lnL_T$

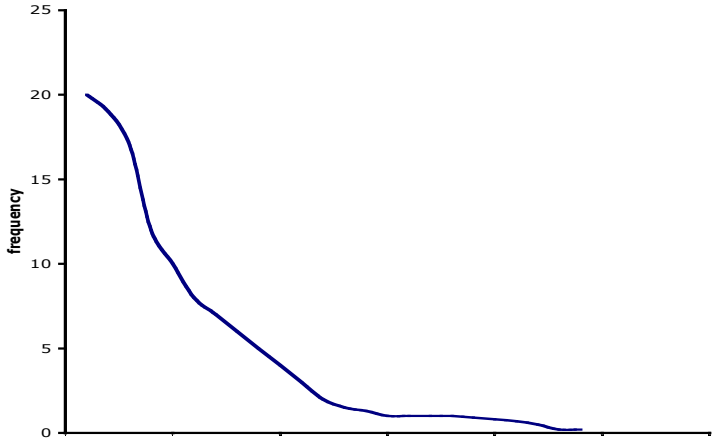
Hypotheses that we wish to test are:

H_0: all trees are equally good explanations of the data

H_1: some or all trees are not equally good explanations of the data

What is the expected distribution of δ under the null?

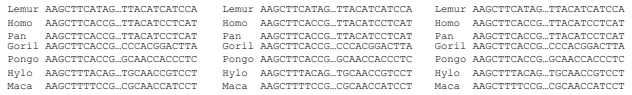
Hint: We know $lnL_{ML} \geq lnL_T$



Shimodaira-Hasegawa (SH-test)

(Shimodaira and Hasegawa, 1999)

• Generate nonparametric bootstrap replicates (could also use RELL)



$\ldots i$

• For each replicate evaluate each candidate tree ($T_{ml}, T_2, T_3, T_4, \ldots T_j$) and center likelihood scores by subtracting the mean *ln*L for that replicate

• Generate Null distribution -- find max adjusted score for each replicate and calculate the difference between Max and each adjusted tree score.

• Test Trees using observed differences -- one-sided test

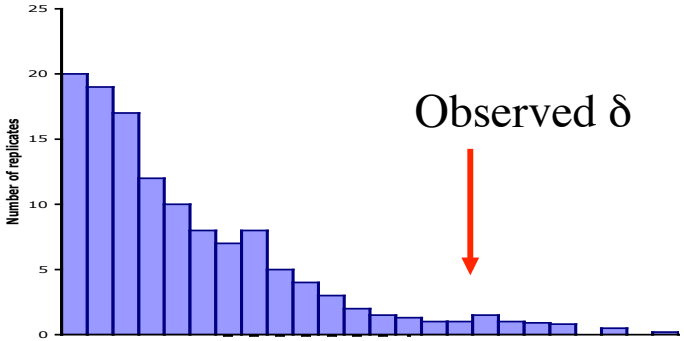
$$(lnL'_1, lnL'_2, lnL'_3, lnL'_4, \ldots lnL'_j)^i$$

$$\propto'_1 = lnL'_1 - \text{mean}(lnL'_1)$$

$$\text{find max}(\propto'_1, \propto'_2, \propto'_3, \propto'_4, \ldots \propto'_j)^I$$

$$\delta'_T = \propto'_{ML} - \propto'_T$$

$$\delta_T = lnL_{ML} - lnL_T \text{ (Observed for each } T_j)$$



$$\delta'_T = \propto'_{ML} - \propto'_T$$

Example with PAUP*

1. Get two trees in memory.
2. Reset PAUP* options to the factory defaults.
3. Compare under the likelihood criterion using the two-tailed KH-test, where the null distribution is based on a normal distribution.
4. Change the test distribution to RELL.
5. Change the test distribution to Full.
6. Do the same for one-tailed KH-test
7. Do the same for the SH-test
8. Do you notice any similarities between the KH-test results and those of the SH-test?

```
reset factory;
hsearch;
lscore 1 2/khtest=normal;
lscore 1 2/khtest=bootstrap;
lscore 1 2/khtest=bootstrap rell=no;
lscore 1 2/khtest=normal tailkh=1;
lscore 1 2/khtest=bootstrap tailkh=1;
lscore 1 2/khtest=bootstrap rell=no tailkh=1;
lscore 1 2/shtest=yes khtest=normal;
lscore 1 2/shtest=yes khtest=bootstrap
 rell=yes;
lscore 1 2/shtest=yes khtest=bootstrap
 rell=no;
```

PAUP* Demonstration

- PAUP* web site:
  - http://paup.sc.fsu.edu/
- Using PAUP*
  - Documentation, command reference/tutorial
    - http://paup.sc.fsu.edu/downl.html
  - Current Protocols in Bioinformatics
    - http://www.currentprotocols.com
      - Chapter 6: Inferring evolutionary relationship

Bibliography

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr., 19:716-723.

Atchley, W. R. and W. M. Fitch. 1991. Gene trees and the origins of inbred strains of mice. Science 254: 554-558.

D. A. Bader, B.M.E Moret, and L. Vawter. 2001. Industrial applications of high-performance computing for phylogeny reconstruction. In H.J. Siegel, (ed.) Proc. SPIE Commercial Applications for High-Performance Computing. Vol. 4428, pp 159-168. Denver, CO SPIE.

Baker and Palumbi 1994, Which whales are hunted? A molecular genetic approach to whaling. Science: 1538-1539.

Bush et al. 1999. Predicting the evolution of human influenza A. Science:1921-1925.

Efron, B. 1985 Bootstrap confidence intervals for a class of parametric problems. Biometrika, 72, 45-58.

Felsenstien, J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Felsenstien. J. 1985. Confidence limits on Phylogenies; an approach using the bootstrap. Evolution 39:783-791.

Felsenstien. J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Genet. 25:471-492.

Fitch, W. M. and W.R. Atchley. 1985. Evolution in inbred strains of mice appears rapid. Science 228:1169-1175.

Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. Science

155:279-284.

Halbur, P., Lum, M. A., Meng, X, Morozov, I., and Paul, P.S. 1994. New porcine reproductive and respiratory syndrome virus DNA and proteins encoded by open reading frames of an Iowa strain of the virus are used in vaccines against PRRSV in pigs. Patent filing WO9606619-A1.

Hillis, D. M. 2000. How to resolve the debate on the origins of AIDS. Science 289:1877-1878.

Hillis, D. M. 2000. Origins of HIV. Science 288:1757-1759.

Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1992. Experimental phylogenetics: generation of a known phylogeny. Science 255:589-592.

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. *Science* 264:671-677.

Holder, M. T. 2001. Using a Complex Model of Sequence Evolution to Evaluate and Improve Phylogenetic Methods. Ph.D. Dissertation. Univ. of Texas at Austin.

Huelsenbeck, J. P., Hillis, D. M. and Jones, R. 1996. Parametric bootstrapping in molecular phylogenetics: Applications and performance. In Ferraris, J. D. and Palumbi, S. R. (eds.), Molecular Zoology. Advances, strategies and protocols. Wiley-Liss, New York, pp. 19-45.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., Bollback, J. P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294: 2310- 2314.

Bibliography (continued)

Goldman, N. 1993. Statistical tests of models of DNA substitution. Journal of Molecular Evolution 36: 182-98.

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Systematic Biology 49:652-670.

Hasegawa, M and H. Kishino. 1989. Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. Evolution 43: 627-677.

Kishino, H. and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. Journal of Molecular Evolution 29:170-179.

Lewis, P. O. 2001. Phylogenetic systematics turns a new leaf . Trends in Evolution and Ecology 16:30-36.

Li, W. 1997. Molecular Evolution. Sinauer Associates. Sunderland, Massachusetts.

Ou, C. et al. 1992. Molecular epidemiology of HIV transmission in a dental practice. 1165-1171

Page, R. D. and Holmes, E. C. 1998. Molecular Evolution: A Phylogenetic Approach. Blackwell Science, Oxford.

Poe, S., and D. L. Swofford. 1999. Taxon sampling revisited. Nature 389:299-300.

Shimodaira, H. and M. Hasegawa. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution 16:1114-1116.

Steel, M. and Penny, D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Molecular Biology and Evolution 17:839-850.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407-514 in D. M. Hillis, C. Moritz, and B. Mable (eds.) *Molecular Systematics* (2nd ed.), Sinauer Associates, Sunderland, Massachusetts.

Yang, Z. 2006. Computational Molecular Evolution. Oxford Unv. Press.