Lecture 6. HMMs for rates. Testing trees, bootstraps, jackknifes[sic]

Joe Felsenstein

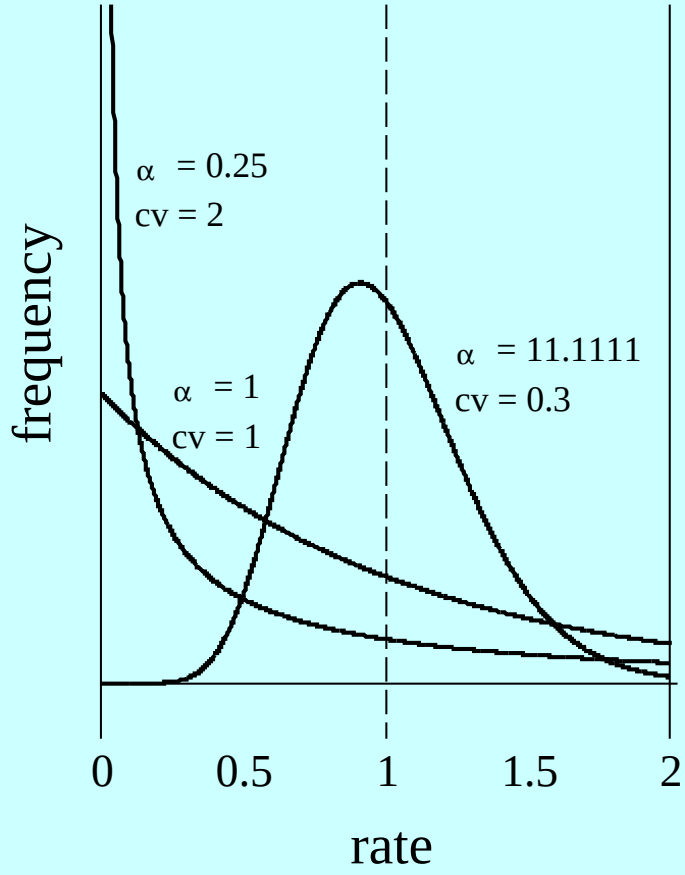
Department of Genome Sciences and Department of Biology

A model of variation in evolutionary rates among sites

The basic idea is that the rate at each site is drawn independently from a distribution of rates. The most widely used choice is the Gamma distribution, which has density function (if its mean is 1):

$$\mathbf{f(r)} \quad = \quad \frac{\alpha^{\alpha}\ \mathbf{r}^{\alpha-1}\ \mathbf{e}^{-\alpha\,\mathbf{r}}}{\mathbf{\Gamma}(\alpha)}$$

Gamma distributions



Gamma distributions with mean 1 and different coefficients of variation (standard deviation / mean). $\alpha = 1/\mathbf{CV^2}$ is the "shape parameter" of the Gamma distribution

Unrealistic aspects of the model:

- There is no reason, aside from mathematical convenience, to assume that the Gamma is the right distribution. A common variation is to assume there is a separate probability $\mathbf{f_0}$ of having rate 0.
- Rates at different sites appear to be correlated, which this model does not allow.
- Rates are not constant throughout evolution – they change with time.

Rates varying among sites

If $\mathbf{L^{(i)}(r_i)}$ is the likelihood of the tree for site $\mathbf{i}$ given that the rate of evolution at site $\mathbf{i}$ is $\mathbf{r_i}$, we can integrate this over a gamma density

$$\mathbf{L^{(i)}} = \int_0^\infty \mathbf{f(r_i;\alpha)L^{(i)}(r_i)\ dr_i}$$

so that the overall likelihood is

$$\mathbf{L} = \prod_{i=1}^{m} \left[\int_0^\infty \mathbf{f(r_i;\alpha)L^{(i)}(r_i)\ dr_i} \right]$$

Unfortunately these integrals cannot be evaluated for trees with more than a few tips as the quantities $\mathbf{L^{(i)}(r_i)}$ complicated.

Hidden Markov Models

These are the most widely used models allowing rate variation to be correlated along the sequence.

We assume:

- There are a finite number of rates, $\mathbf{m}$. Rate i is $\mathbf{r_i}$.
- There are probabilities $\mathbf{p_i}$ of a site having rate $\mathbf{i}$.
- A process not visible to us ("hidden") assigns rates to sites. It is a Markov process working along the sequence. For example it might have transition probability $\mathrm{Prob}\,(\mathbf{j}|\mathbf{i})$ of changing to rate $\mathbf{j}$ in the next site, given that it is at rate $\mathbf{i}$ in this site.
- The probability of our seeing some data are to be obtained by summing over all possible combinations of rates, weighting approproately by their probabilities of occurrence.

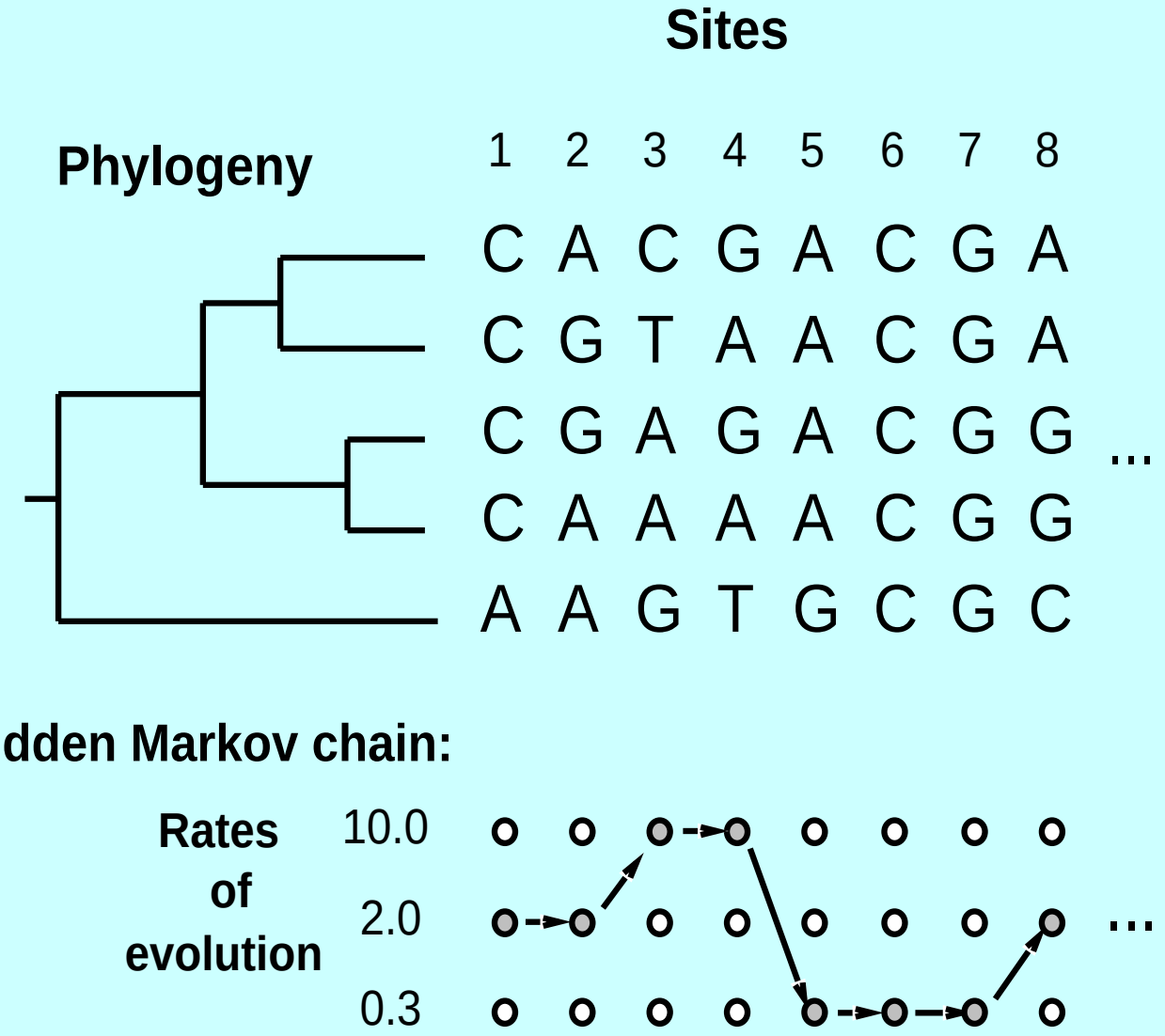
Likelihood with a[n] HMM

Suppose that we have a way of calculating, for each possible rate at each possible site, the probability of the data at that site given that rate. This is

$$\text{Prob}\left(\mathbf{D^{(i)}} \mid \mathbf{r_j}\right)$$

To get the overall probability of all data, sum over all possible paths through the array of sites $\times$ rates, weighting each combination of rates by its probability:

A Hidden Markov Model for rates in a phylogeny



Hidden Markov Models

If there are a number of hidden rate states, with state i having rate r_i

$$\text{Prob}\,(\mathbf{D} \mid \mathbf{T}) = \sum_{\mathbf{i_1}} \sum_{\mathbf{i_2}} \ldots \sum_{\mathbf{i_p}} \text{Prob}\,(\mathbf{r_{i_1}}, \mathbf{r_{i_2}}, \ldots \mathbf{r_{i_p}})$$

$$\times\, \text{Prob}\,(\mathbf{D} \mid \mathbf{T}, \mathbf{r_{i_1}}, \mathbf{r_{i_2}}, \ldots \mathbf{r_{i_m}})$$

Evolution is independent once each site has had its rate specified

$$\text{Prob}\,(\mathbf{D} \mid \mathbf{T}, \mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r_p}) \;=$$

$$\prod_{\mathbf{i=1}}^{\mathbf{p}} \text{Prob}\,(\mathbf{D^{(i)}} \mid \mathbf{T}, \mathbf{r_i}).$$

Seems impossible ...

Evolution is independent once each site has had its rate specified

$$\text{Prob}\,(\mathbf{D}|\mathbf{T},\mathbf{r_1},\mathbf{r_2},\ldots,\mathbf{r_m}) \;=\; \prod_{\mathbf{i=1}}^{\mathbf{m}} \text{Prob}\,(\mathbf{D^{(i)}}|\mathbf{T},\mathbf{r_i}).$$

To compute the likelihood we sum over all ways rate states could be assigned to sites:

$$\begin{aligned}
\mathbf{L} &= \text{Prob}\,(\mathbf{D}\mid\mathbf{T}) \\
&= \sum_{\mathbf{i_1=1}}^{\mathbf{m}} \sum_{\mathbf{i_2=1}}^{\mathbf{m}} \ldots \sum_{\mathbf{i_p=1}}^{\mathbf{m}} \text{Prob}\,\left(\mathbf{r_{i_1}},\mathbf{r_{i_2}},\ldots,\mathbf{r_{i_p}}\right) \\
&\quad \times \text{Prob}\,\left(\mathbf{D^{(1)}}\mid\mathbf{r_{i_1}}\right)\text{Prob}\,\left(\mathbf{D^{(2)}}\mid\mathbf{r_{i_2}}\right)\ldots\text{Prob}\,\left(\mathbf{D^{(n)}}\mid\mathbf{r_{i_p}}\right)
\end{aligned}$$

Problem: The number of rate combinations is very large. With 100 sites and 3 rates at each, it is $\mathbf{3^{100} \simeq 5 \times 10^{47}}$. This makes the summation impractical.

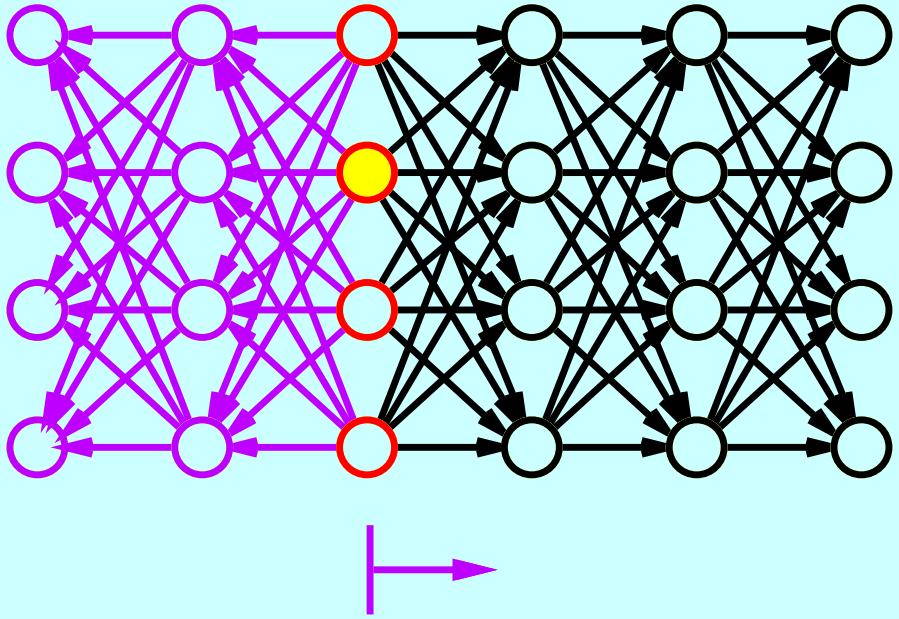
This is an HMM

The hidden states identify the rates that applied at a site. Each rate implies (together with the tree, which is in common to all sites) a distribution of possible base patterns (4^n of them if there are n sequences on the tree). At each site one has actually occurred.

We can use the usual Forwards Algorithm to sum up likelihood over all paths through the array of rates.

The forwards-backwards algorithm.

the "forward−backward" algorithm allows us to get
the probability of everything given one site's state



... which enables us to compute the fraction of the likelihood contributed by one of the rates at one of the sites. Alternatively, the Viterbi algorithm enables us to find the single path that contributes the most to the likelihood.

PhyloHMMs

Siepel and Haussler (2004) have called the HMMs over rates (and other HMMs that operate along multiple sequence alignments and evaluate likelihoods on a tree) “phylo-HMMs”. They have applied these widely to search for conserved regions in alignments of genomes and for gene-finding.

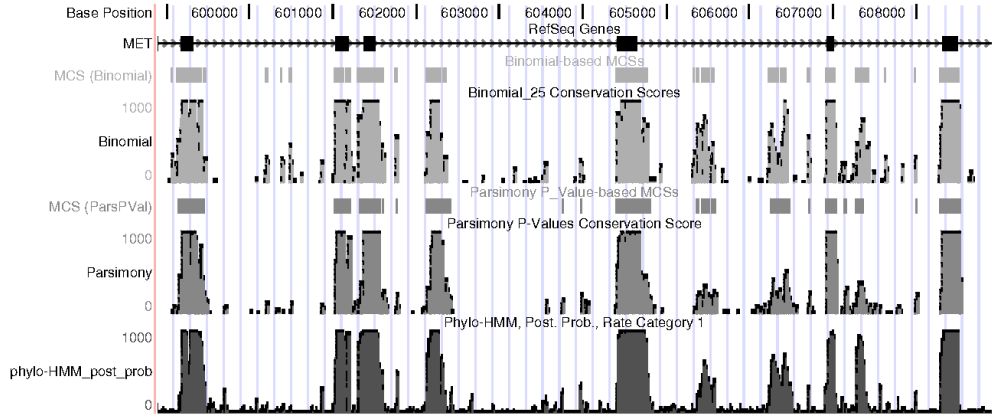
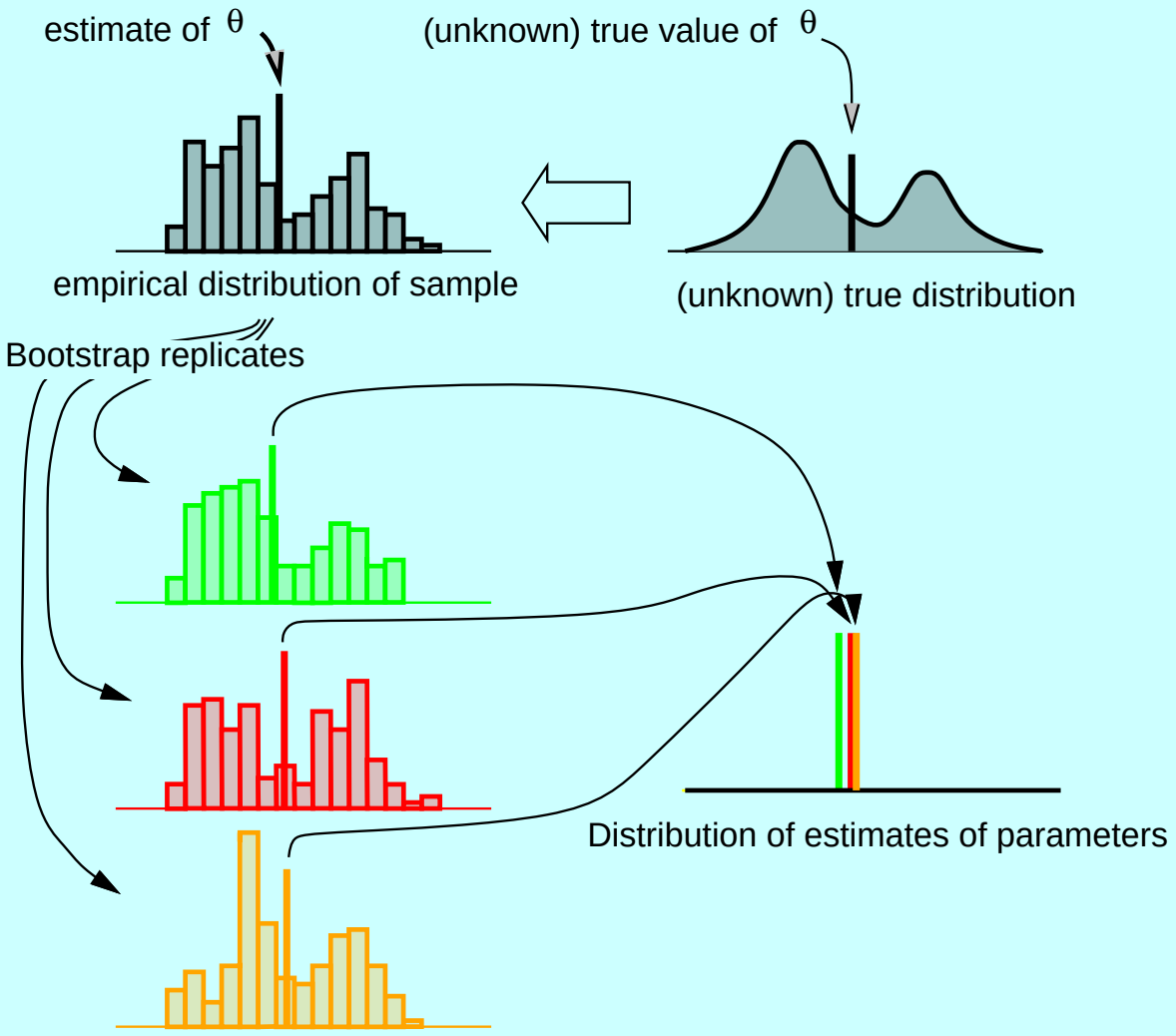


Fig. 5. A screen shot from the UCSC Genome Browser [24] showing a selected region of the data set of example 2, including several exons of the *MET* gene (black boxes at top). The binomial-based (light gray) and parsimony-based (medium gray) conservation scores of Margulies et al. [30] are shown as tracks in the browser, as are the posterior probabilities ($\times 1000$) of state s_1 in the phylo-HMM (dark gray). Plots similar to this one, showing phylo-HMM-based conservation scores across the whole human genome, can be viewed online at http://genome.ucsc.edu.

A non-phylogeny example of the bootstrap



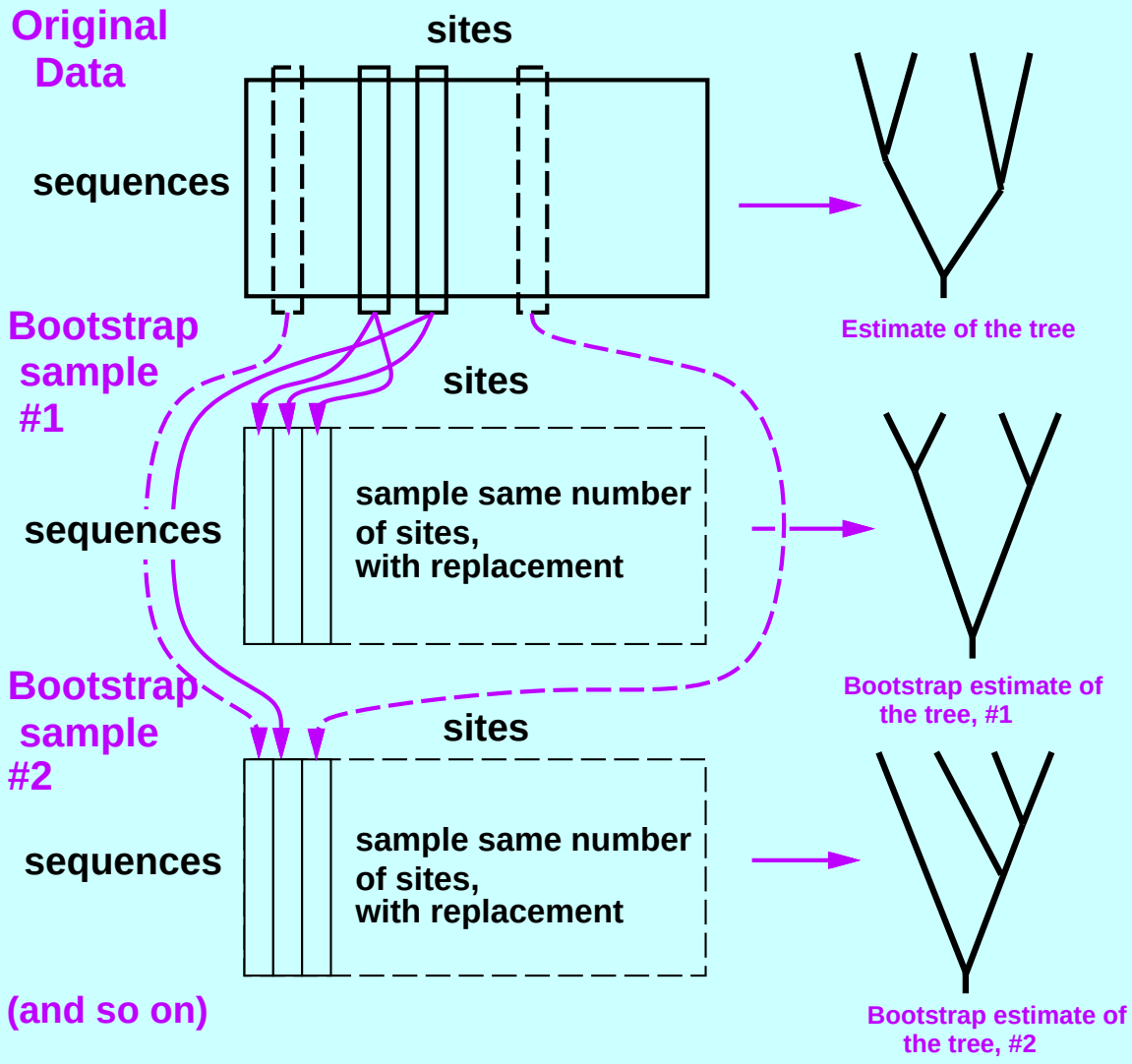
Bootstrap sampling from a distribution (a mixture of two normals) to estimate the variance of the mean

Bootstrap sampling

To infer the error in a quantity, θ, estimated from a sample of points $x_1, x_2, \ldots, x_n$ we can

- Do the following R times ($R = 1000$ or so)
- Draw a “bootstrap sample" by sampling n times with replacement from the sample. Call these $x_1^*, x_2^*, \ldots, x_n^*$. Note that some of the original points are represented more than once in the bootstrap sample, some once, some not at all.
- Estimate θ from the bootstrap sample, call this $\hat{\theta}_k^*$ ($k = 1, 2, \ldots, R$)
- When all R bootstrap samples have been done, the distribution of $\hat{\theta}_i^*$ estimates the distribution one would get if one were able to draw repeated samples of n points from the unknown true distribution.

Bootstrap sampling of phylogenies

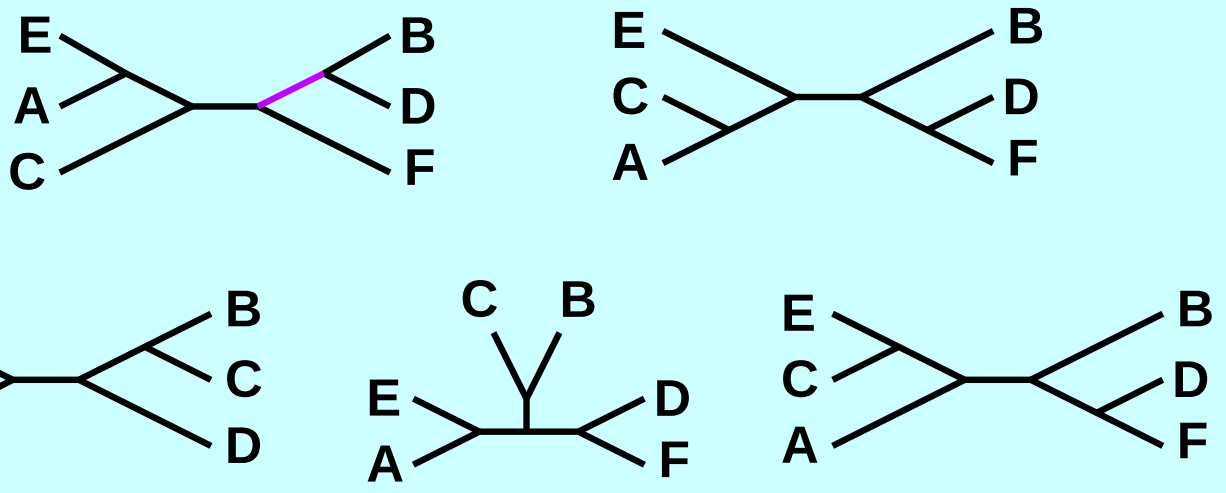


More on the bootstrap for phylogenies

- The sites are assumed to have evolved independently given the tree. They are the entities that are sampled (the x_i).
- The trees play the role of the parameter. One ends up with a cloud of R sampled trees.
- To summarize this cloud, we ask, for each branch in the tree, how frequently it appears among the cloud of trees.
- We make a tree that summarizes this for all the most frequently occurring branches.
- This is the *majority rule consensus tree* of the bootstrap estimates of the tree.

A partition on the first tree

Trees:

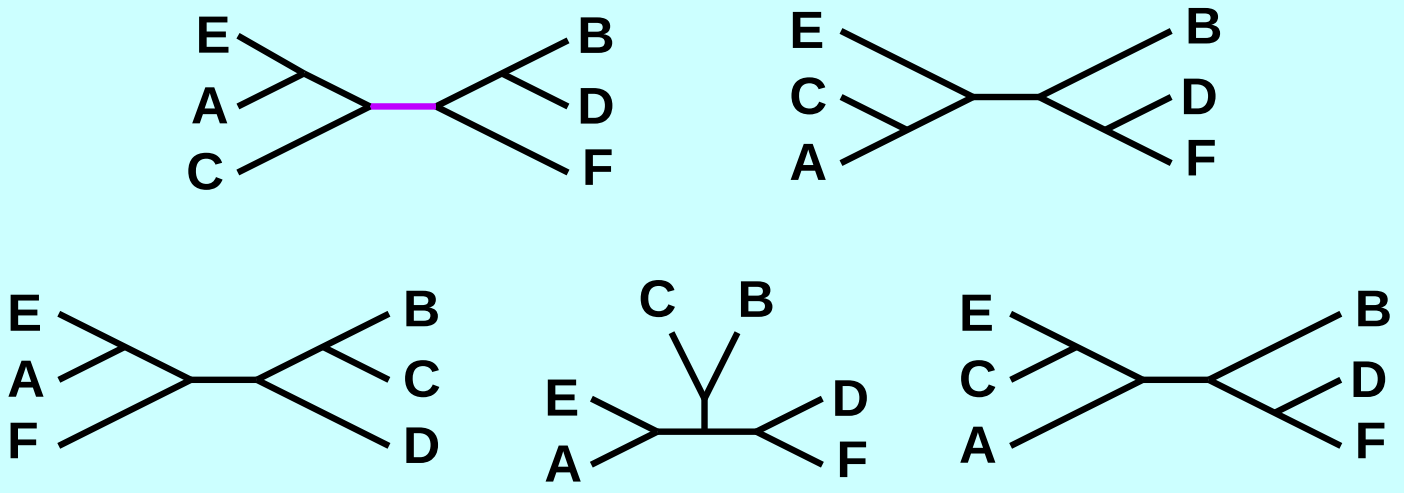


How many times each partition of species is found:

AE | BCDF
ACE | BDF
ACEF | BD **1**
AC | BDEF
AEF | BCD
ADEF | BC
ABDF | EC
ABCE | DF

A second partition

Trees:

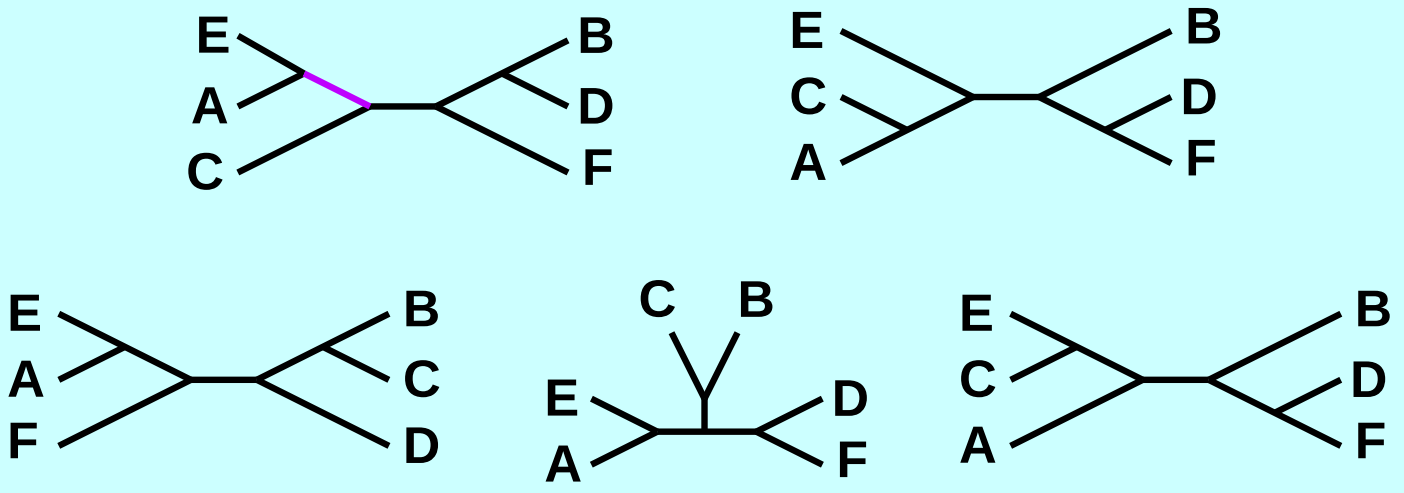


How many times each partition of species is found:

AE | BCDF	
ACE | BDF	**1**
ACEF | BD	**1**
AC | BDEF	
AEF | BCD	
ADEF | BC	
ABDF | EC	
ABCE | DF	

A third partition

Trees:

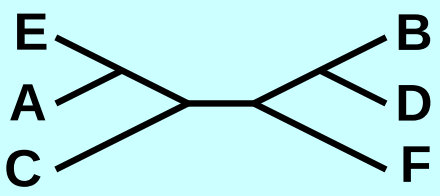


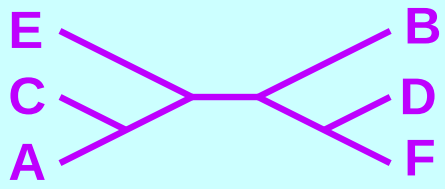
How many times each partition of species is found:

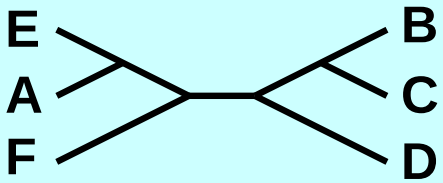
AE | BCDF	**1**
ACE | BDF	**1**
ACEF | BD	**1**
AC | BDEF	
AEF | BCD	
ADEF | BC	
ABDF | EC	
ABCE | DF	

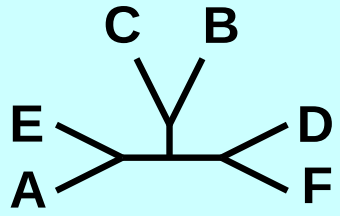
The second tree and its partitions

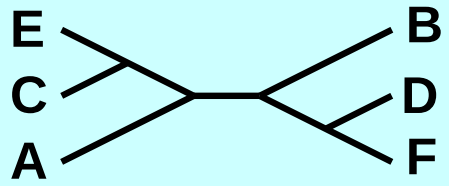
Trees:









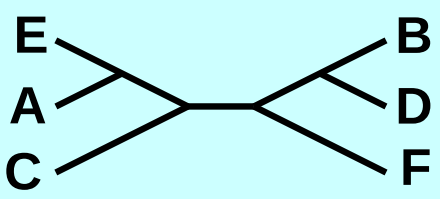


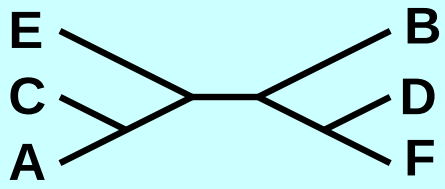
How many times each partition of species is found:

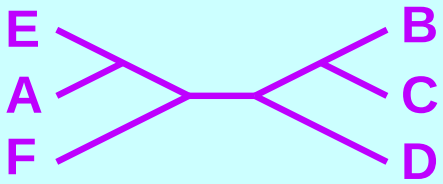
AE | BCDF	**1**
ACE | BDF	**2**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	
ADEF | BC	
ABDF | EC	
ABCE | DF	**1**

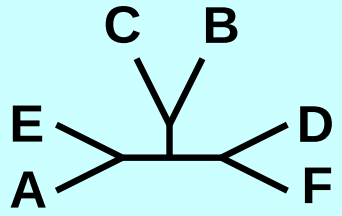
The third tree and its partitions

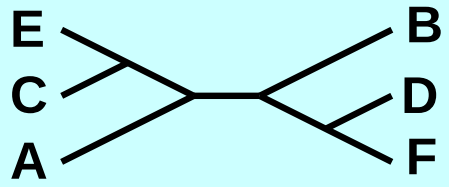
Trees:









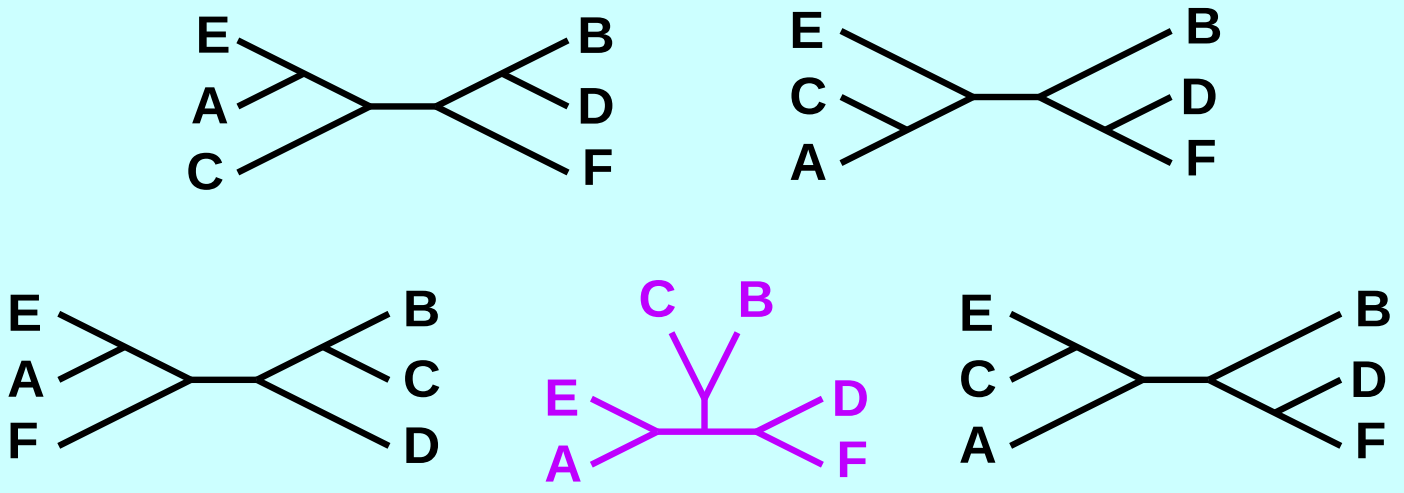


How many times each partition of species is found:

AE | BCDF	**2**
ACE | BDF	**2**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**1**
ABDF | EC	
ABCE | DF	**1**

The fourth tree and its partitions

Trees:

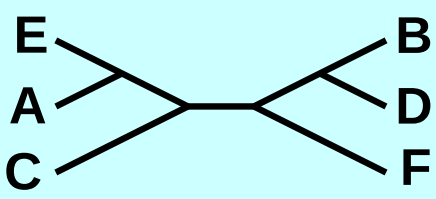


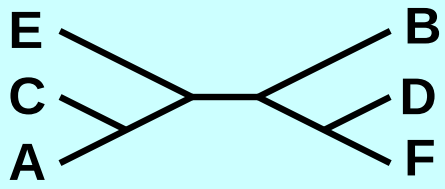
How many times each partition of species is found:

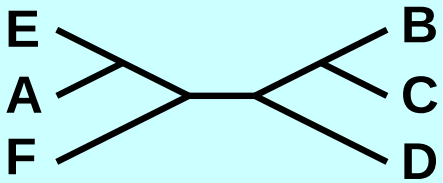
AE | BCDF	**3**
ACE | BDF	**2**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**2**
ABDF | EC	
ABCE | DF	**2**

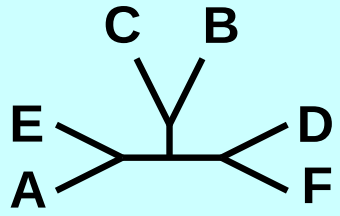
The fifth tree and its partitions

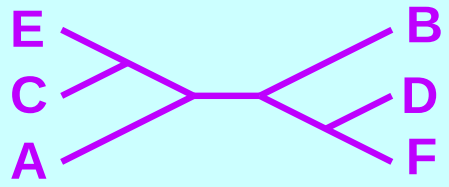
Trees:









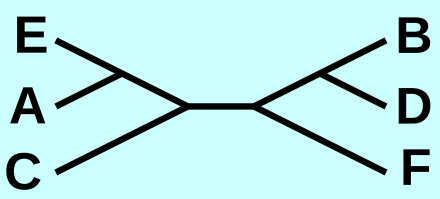


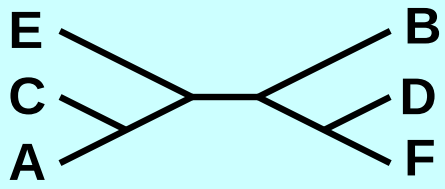
How many times each partition of species is found:

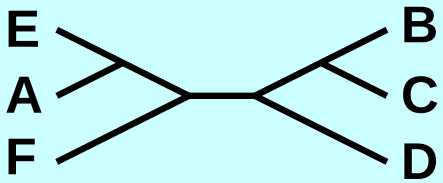
AE | BCDF	**3**
ACE | BDF	**3**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**2**
ABDF | EC	**1**
ABCE | DF	**3**

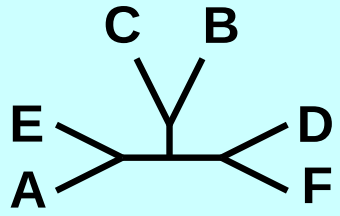
The trees and their partitions

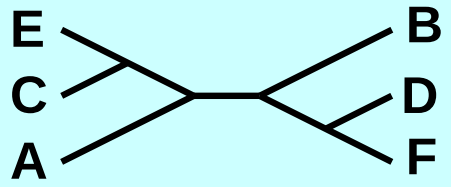
Trees:









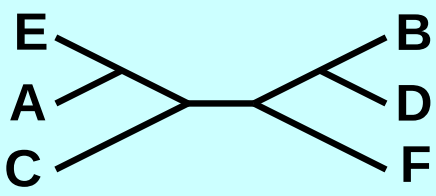


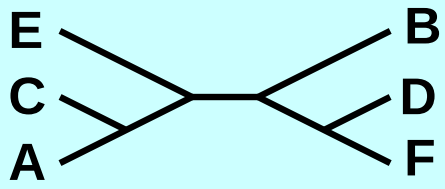
How many times each partition of species is found:

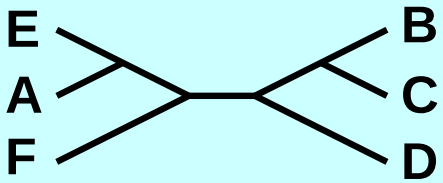
AE | BCDF	**3**
ACE | BDF	**3**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**2**
ABDF | EC	**1**
ABCE | DF	**3**

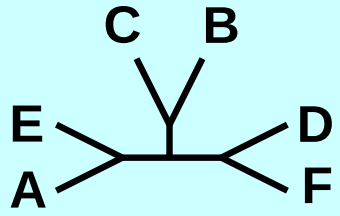
Majority rule consensus trees

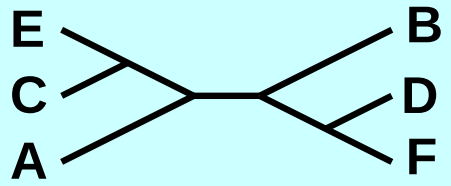
Trees:





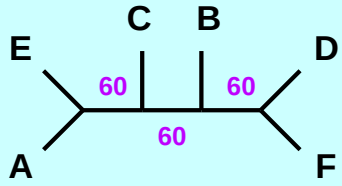






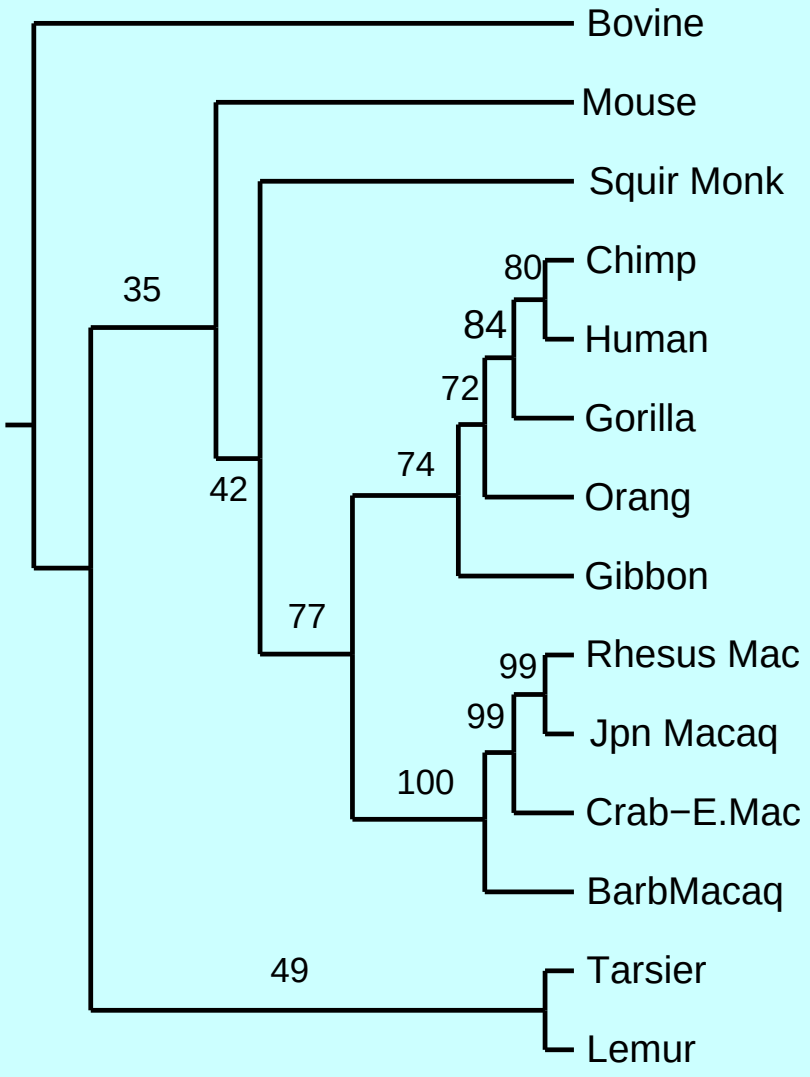
How many times each partition of species is found:

AE | BCDF	**3**
ACE | BDF	**3**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**2**
ABDF | EC	**1**
ABCE | DF	**3**



Majority−rule consensus tree of the unrooted trees:

An example of bootstrap sampling of trees



232 nucleotide, 14-species mitochondrial D-loop analyzed by parsimony, 100 bootstrap replicates

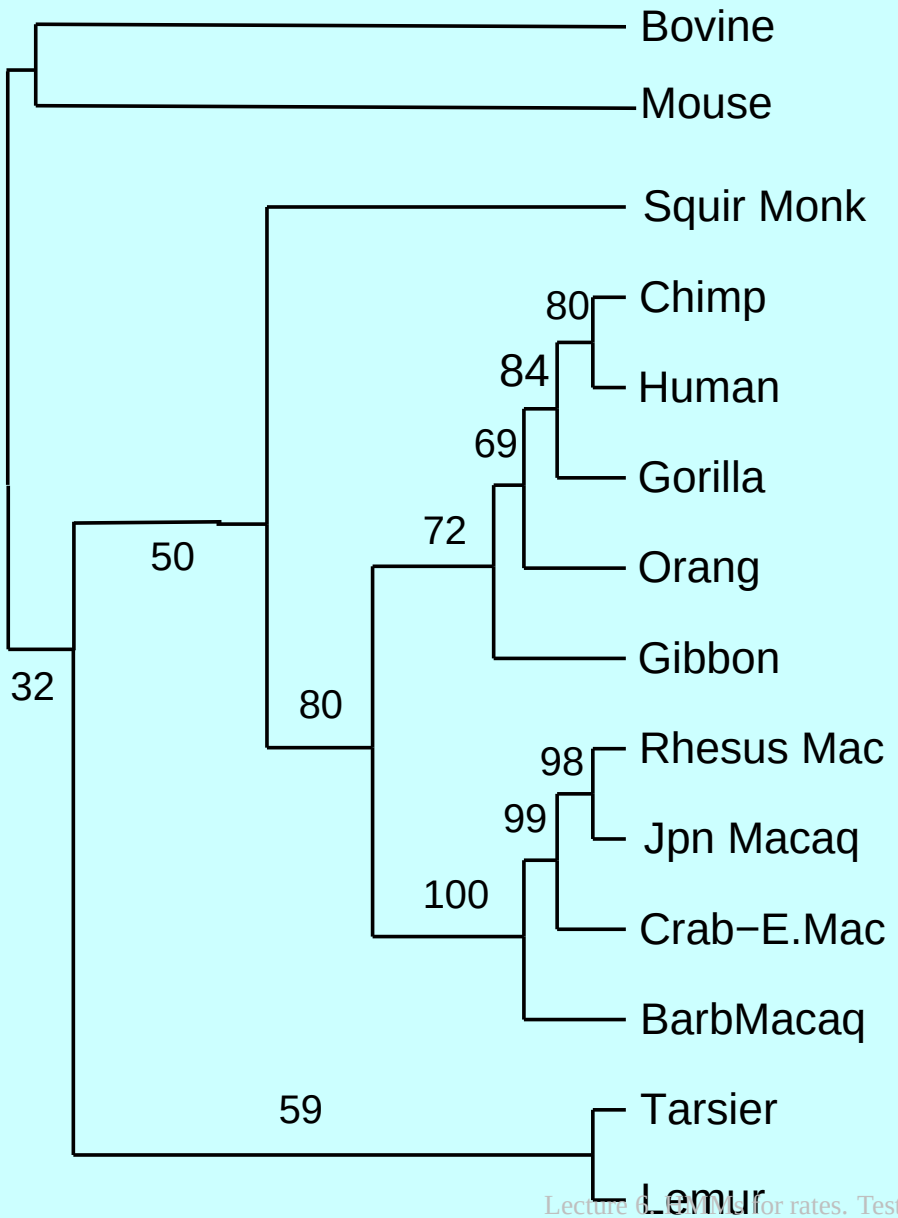
Potential problems with the bootstrap

1. Sites may not evolve independently
2. Sites may not come from a common distribution (but can consider them sampled from a mixture of possible distributions)
3. If do not know which branch is of interest at the outset, a “multiple-tests" problem means P values are overstated
4. P values are biased (too conservative)
5. Bootstrapping does not correct biases in phylogeny methods

Other resampling methods

- Delete-half jackknife. Sample a random 50% of the sites, *without* replacement.
- Delete-$1/e$ jackknife (Farris et. al. 1996) (too little deletion from a statistical viewpoint).
- Reweighting characters by choosing weights from an exponential distribution.
- In fact, reweighting them by any exchangeable weights having coefficient of variation of 1
- Parametric bootstrap – simulate data sets of this size assuming the estimate of the tree is the truth
- (to correct for correlation among adjacent sites) (Künsch, 1989) Block-bootstrapping – sample n/b blocks of b adjacent sites.

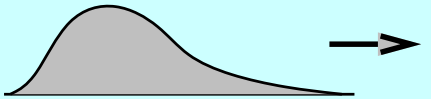
Delete half jackknife on the example



Parametric bootstrapping

The Parametric Bootstrap (Efron, 1985)

Suppose we have independent observations drawn from a known distribution:

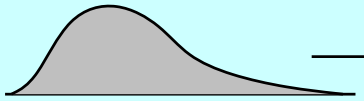


$x_1, x_2, x_3, \ldots x_n \rightarrow \theta$

and a parameter, θ, calculated from this.

To infer the variability of θ
Use the current estimate, $\hat{\theta}$
Use the distribution that has that as its true parameter

and take the distribution of the $\hat{\theta}_i$ as the estimate of the distribution from which it is drawn



sample R data sets from that distribution, each having the same sample size as the original sample

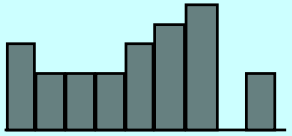
$x_1^*, x_2^*, x_3^*, \ldots x_n^* \rightarrow \hat{\theta}_1$

$x_1^*, x_2^*, x_3^*, \ldots x_n^* \rightarrow \hat{\theta}_2$

$x_1^*, x_2^*, x_3^*, \ldots x_n^* \rightarrow \hat{\theta}_3$

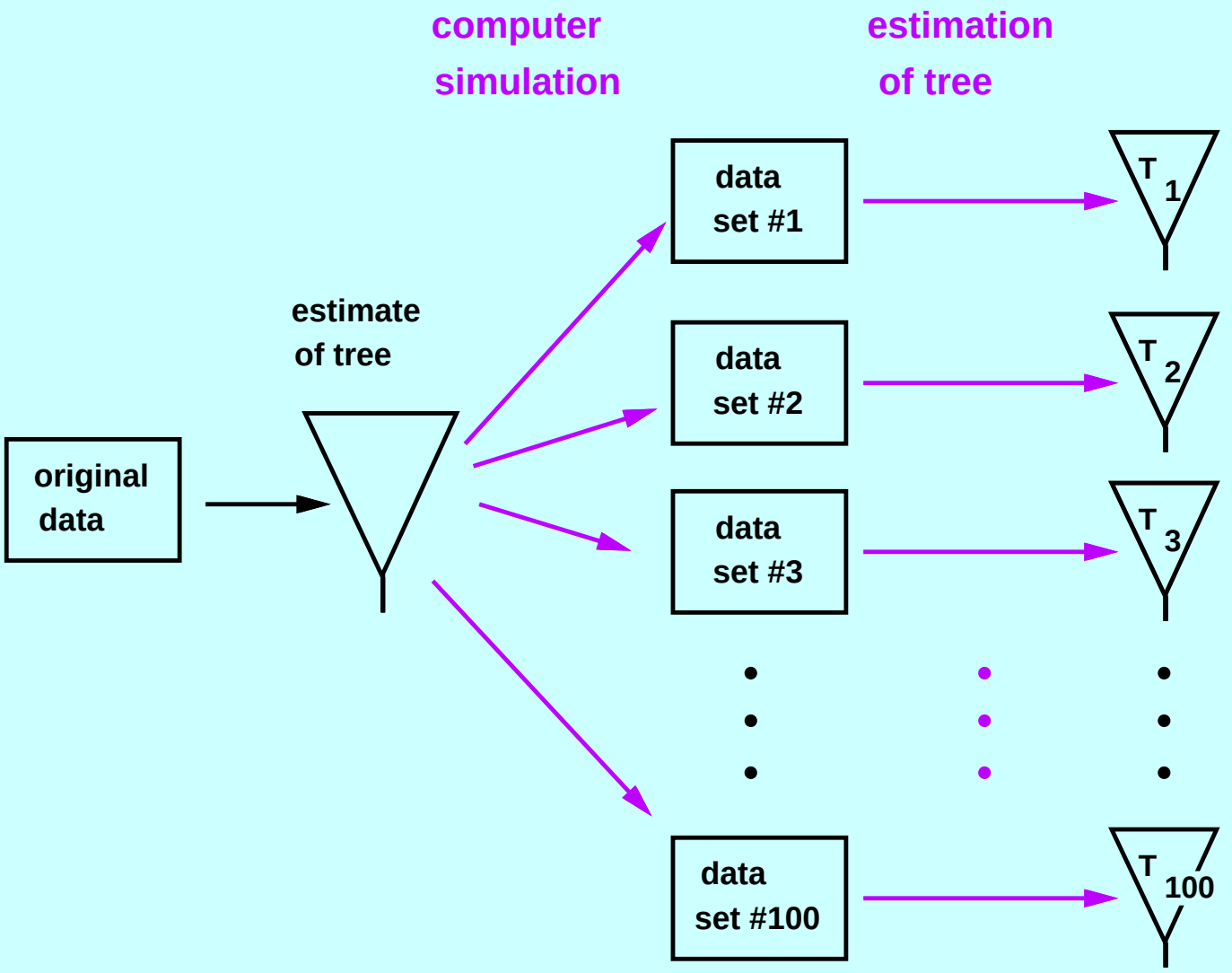
$\vdots$

$x_1^*, x_2^*, x_3^*, \ldots x_n^* \rightarrow \hat{\theta}_R$

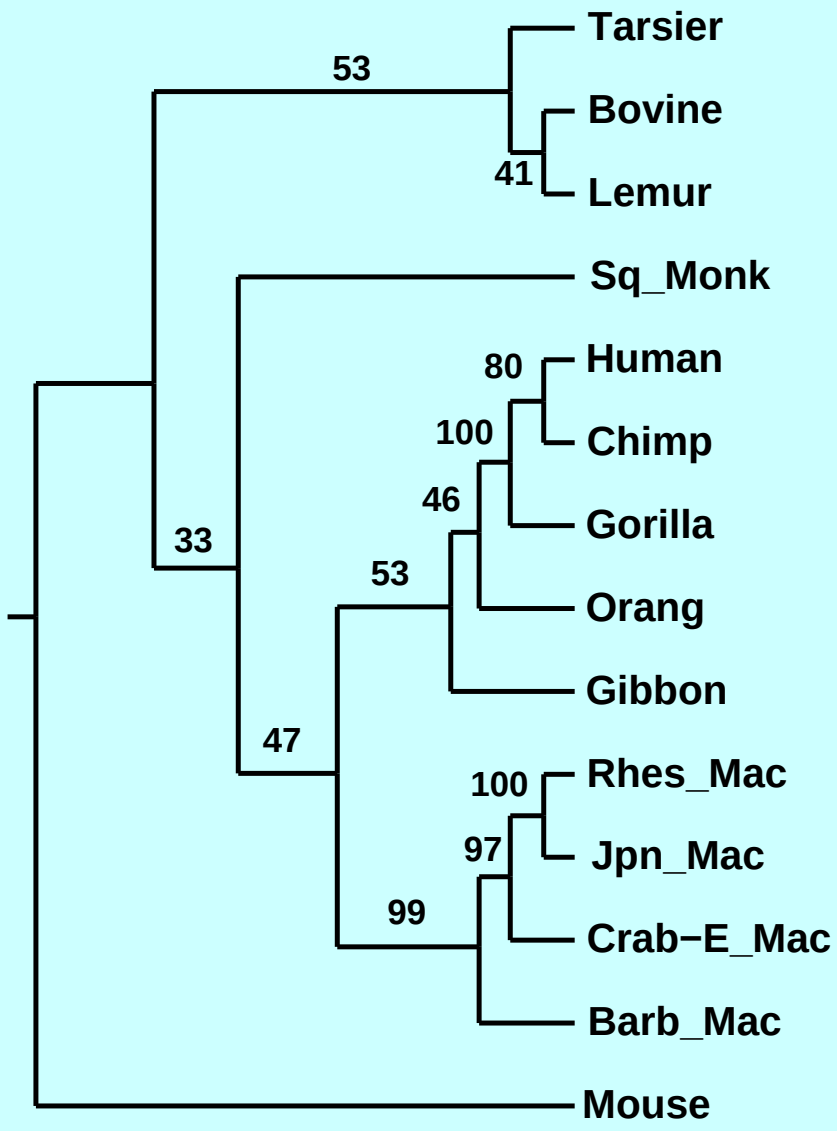


θ

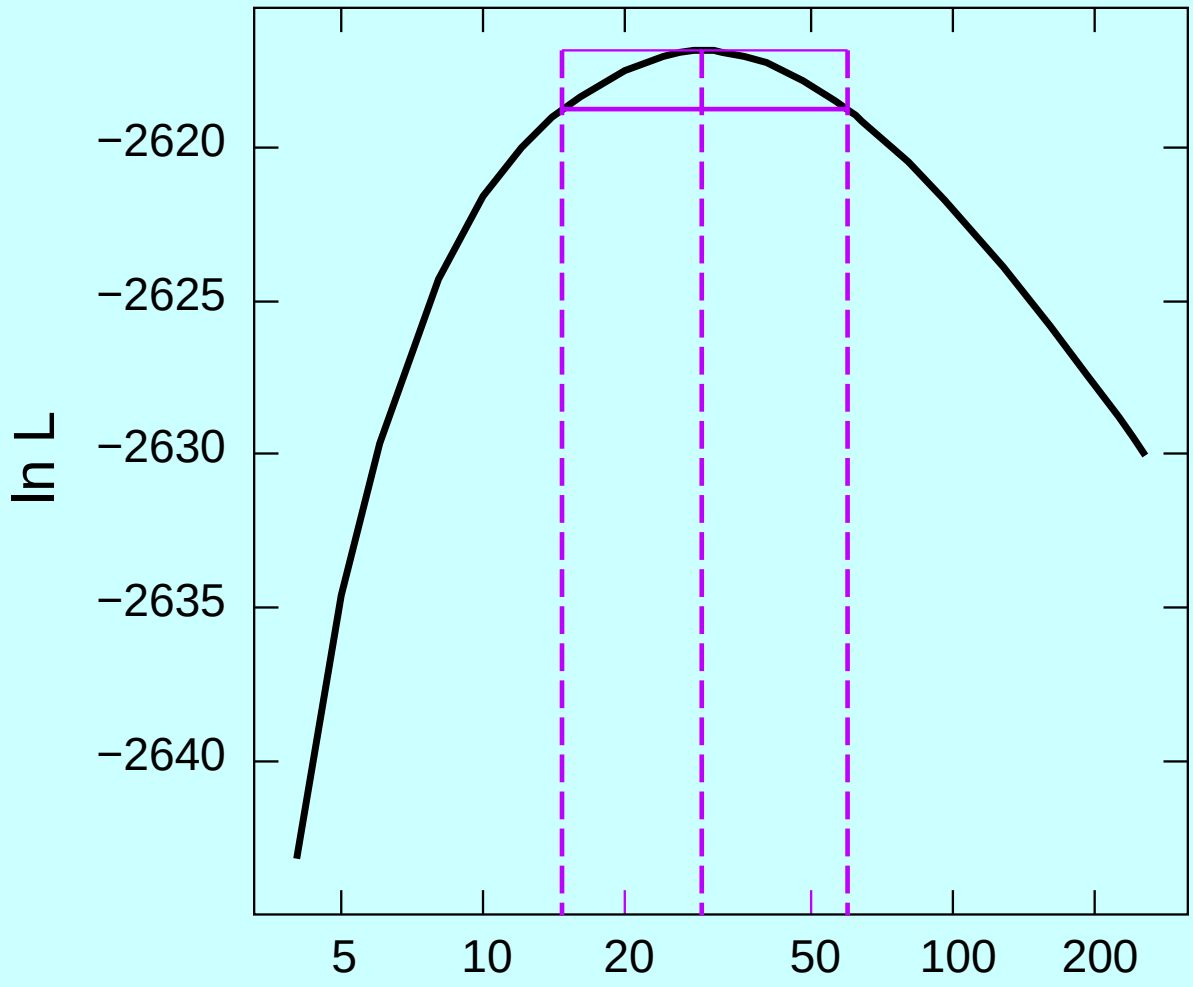
The parametric bootstrap for phylogenies



An example of the parametric bootstrap

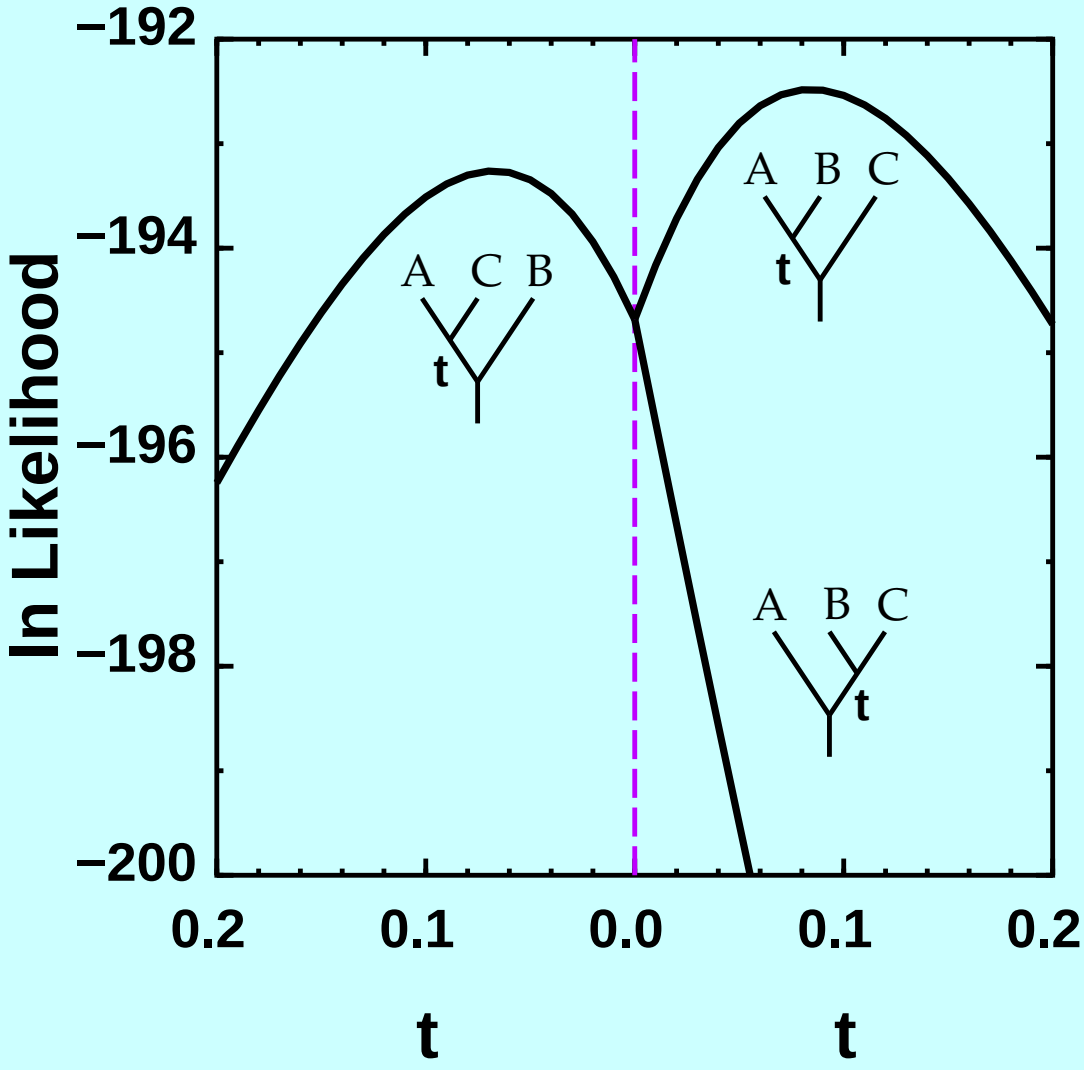


Likelihood ratio confidence limits on Ts/Tn ratio

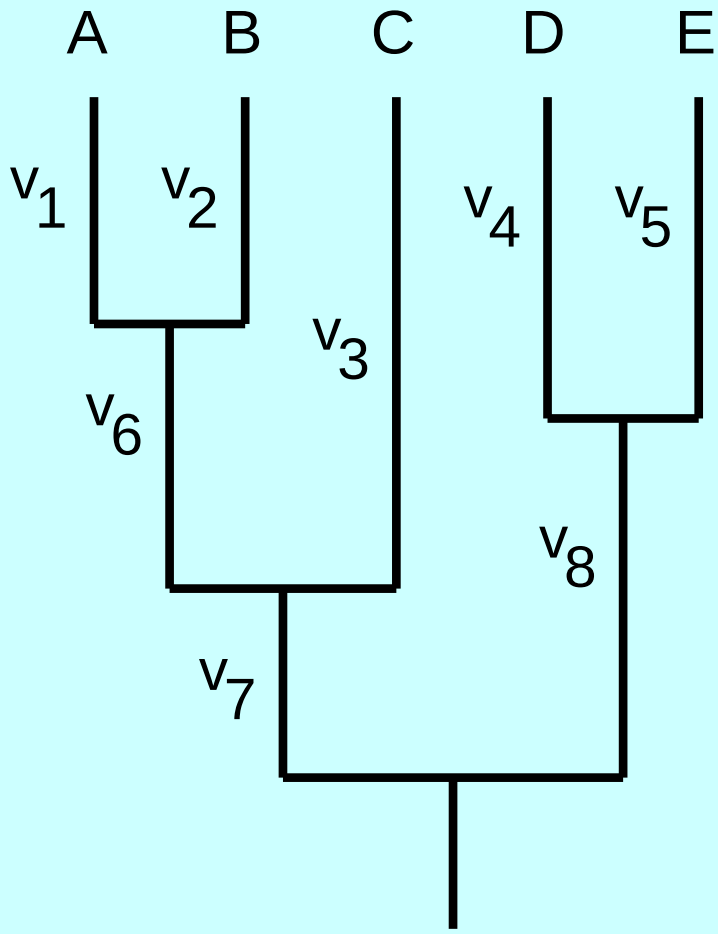


for the 14-species primate data set

Likelihoods in tree space – a 3-species clock example



The constraints for a molecular clock



Constraints for a clock

$$v_1 = v_2$$

$$v_4 = v_5$$

$$v_1 + v_6 = v_3$$

$$v_3 + v_7 = v_4 + v_8$$

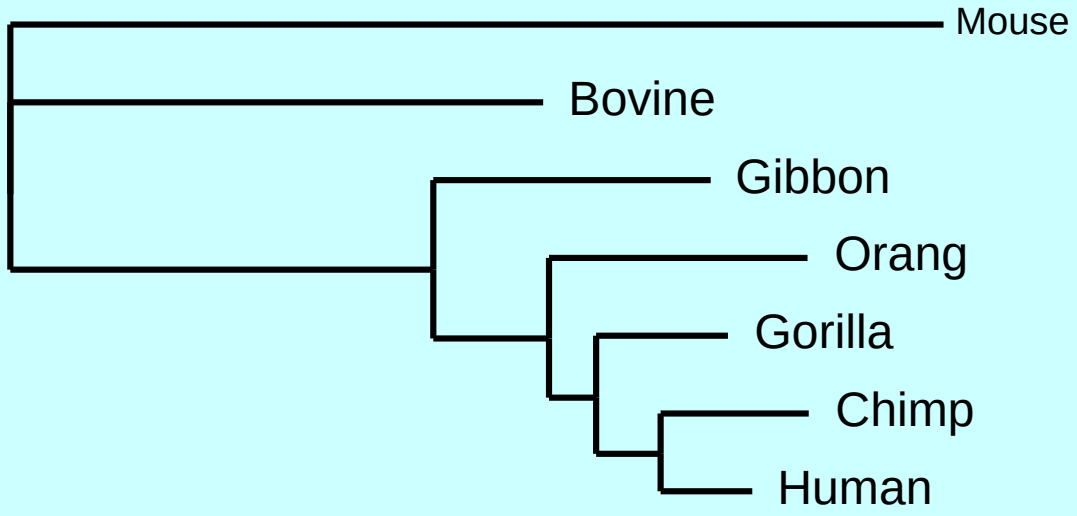
Testing for a molecular clock

To test for a molecular clock:

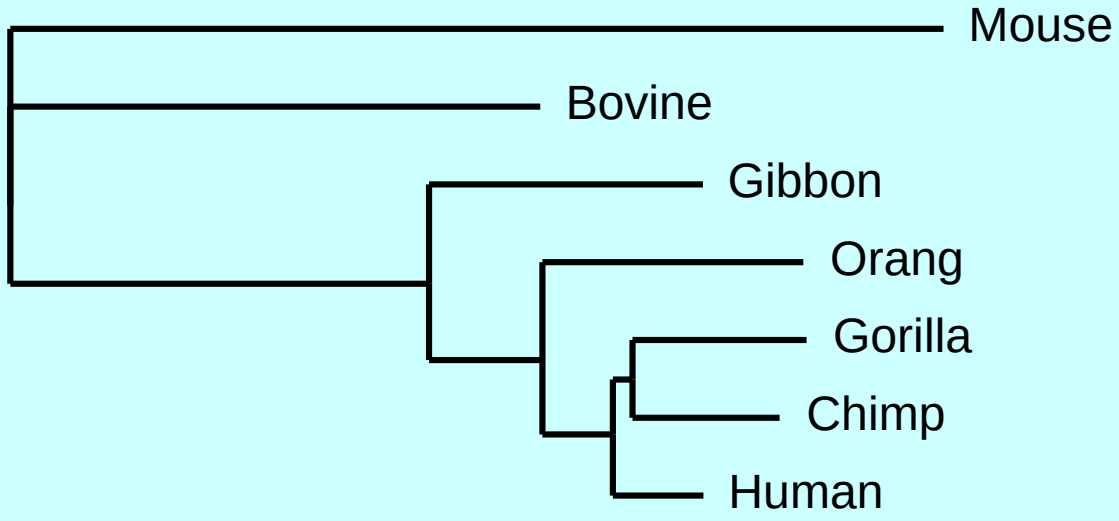
- Obtain the likelihood with no constraint of a molecular clock (For primates data with $T_s/T_n = 30$ we get $\ln L_1 = -2616.86$
- Obtain the highest likelihood for a tree which is constrained to have a molecular clock: $\ln L_0 = -2679.0$
- Look up $2(\ln L_1 - \ln L_0) = 2 \times 62.14 = 124.28$ on a χ^2 distribution with $n - 2 = 12$ degrees of freedom (in this case the result is significant)

Two trees to be tested by paired sites tests

Tree I



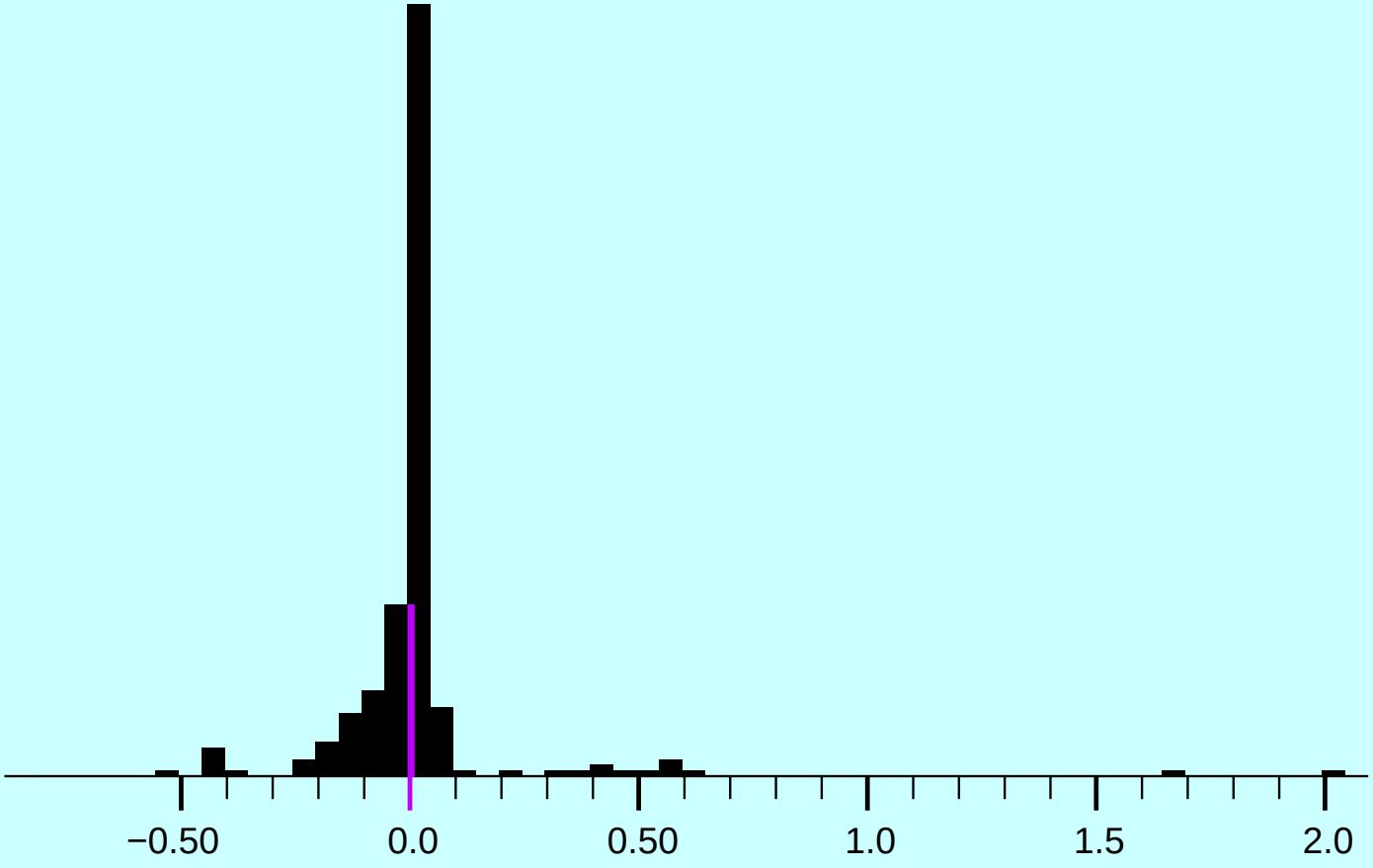
Tree II



Differences in log likelihoods site by site

Tree \ site	1	2	3	4	5	6	…	231	232	ln L
I	−2.971	−4.483	−5.673	−5.883	−2.691	−8.003	…	−2.971	−2.691	−1405.61
II	−2.983	−4.494	−5.685	−5.898	−2.700	−7.572	…	−2.987	−2.705	−1408.80
Diff	+0.012	+0.111	+0.013	+0.015	+0.010	−0.431	…	+0.012	+0.010	+3.19

Histogram of log likelihood differences



Paired sites tests

- Winning sites test (Prager and Wilson, 1988). Do a sign test on the signs of the differences.
- z test (me, 1993 in PHYLIP documentation). Assume differences are normal, do z test of whether mean (hence sum) difference is significant.
- t test. Swofford et. al., 1996: do a t test (paired)
- Wilcoxon ranked sums test (Templeton, 1983).
- RELL test (Kishino and Hasegawa, 1989 per my suggestion). Bootstrap resample sites, get distribution of difference of totals.

In this example

- Winning sites test. 160 of 232 sites favor tree I. $P < 3.279 \times 10^{-9}$
- z test. Difference of log-likeihood totals is 0.948104 standard deviations from 0, $P = 0.343077$. Not significant.
- `t` test. Same as `z` test for this large a number of sites.
- Wilcoxon ranked sums test. Rank sum is 4.82805 standard deviations below its expected value, $P = 0.000001378765$
- RELL test. 8,326 out of 10,000 samples have a positive sum, $P = 0.3348$ (two-sided)

Bayesian methods

Bayesian methods have gained popularity, some of it because they can be computationally faster than bootstrapping (in my view you should use them if you agree with them, not just because of speed).

In the Bayesian framework, one can avoid the separate calculation of confidence intervals. The posterior distribution of trees shows us how much credence to give different trees (for example, it assigns probabilities to different tree topologies).

The interesting issue is how to summarize this posterior distribution in the best way. In this respect Bayesian methods leave you in a situation analogous to having the cloud of bootstrap-sampled trees without yet having summarized them.

Clade probabilities, computed in the same way as bootstrap probabilities from the posterior cloud of trees, are a popular way of summarizing this. They are used in the popular Bayesian program MrBayes.

References, page 1

Jin, L. and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution* **7:** 82-102. [Gamma distributed rates in a distance]

Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symposia on Quantitative Biology* **52:** 825-837. [Lognormal rates in a distance]

Waddell, P. J. and M. A. Steel. 1997. General time-reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites. *Molecular Phylogenies and Evolution* **8:** 398-414. [More generalized rate distributions in a more generalized distance]

Churchill, G.A. 1989. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51:** 79-94. [First paper using HMMs on sequences, without trees]

Felsenstein, J. and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution *Molecular Biology and Evolution* **13:** 93-104. [HMMs for rates in ML trees]

References, page 2

Siepel, A., and D. Haussler. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology* **11(2-3):** 413-428. [Generalizes the use of "phylo-HMMs" and applies them to searching for conserved regions in genomes]

Yang, Z. 1994. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10:** 1396-1401. [Rates varying in gamma distribution in an ML tree method for few species]

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39:** 306-314. [First paper on HMMs for ML trees, used to approximate Gamma distributions for more species]

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* **139:** 993-1005. [Also allowing for correlated rates along the molecule]

References

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7:** 1-26. [The original bootstrap paper]

Margush, T. and F. R. McMorris. 1981. Consensus n-trees. *Bulletin of Mathematical Biology* **43:** 239-244i. [Majority-rule consensus trees]

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39:** 783-791. [The bootstrap first applied to phylogenies]

Farris, J. S., V. A. Albert, M. Kallersjö, D. Lipscomb, and A. G. Kluge. 1996. Parsimony jackknifing outperforms neighbor-joining. *Cladistics* **12:** 99-124. [The delete-$1/e$ jackknife for phylogenies]

Wu, C. F. J. 1986. Jackknife, bootstrap and other resampling plans in regression analysis. *Annals of Statistics* **14:** 1261-1295. [The delete-half jackknife]

bias in the bootstrap

Zharkikh, A., and W.-H. Li. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution* **9:** 1119-1147. [Discovery and explanation of bias in P values]

Hillis, D. M. and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42:** 182-192. [Bias in P values seen in a large simulation study]

Felsenstein, J. and H. Kishino. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology* **42:** 193-200. [A more detailed exposition of the bias of P values in a normal case]

Sanderson, M. J. 1995. Objections to bootstrapping phylogenies: a critique. *Systematic Biology* **44:** 299-320. [Good but he accepts a few criticisms I would not have accepted]

variations on the bootstrap

Harshman, J. 1994. The effect of irrelevant characters on bootstrap values. *Systematic Zoology* **43:** 419-424. [Not much effect on bootstrap support with parsimony whether or not you include invariant characters]

Künsch, H. R. 1989. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17:** 1217-1241. [The block-bootstrap]

Efron, B. 1985. Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72:** 45-58. [The parametric bootstrap]

Templeton, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* **37:** 221-244. [First paper on KHT test]

Prager, E. M. and A. C. Wilson. 1988. Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. *Journal of Molecular Evolution* **27:** 326-335. [winning-sites test]

paired sites tests

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. [The KHT test]

Hasegawa, M., H. Kishino. 1989. Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution* **43:** 672-677 [The KHT test]

Hasegawa, M. and H. Kishino. 1994. Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree. Molecular Biology and Evolution **11:** 142-145. [RELL probabilities]

Shimodaira, H. and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16:** 1114-1116. [SH test for multiple trees]

paired sites, and miscellaneous

Cavender, J. A. 1977. Taxonomy with confidence. *Mathematical Biosciences* **40:** 271-280 (Erratum, vol. 44, p. 308, 1979) [First paper on testing trees]

Felsenstein, J. 1985c. Confidence limits on phylogenies with a molecular clock. *Systematic Zoology* **34**: 152-161. [A 3-species case where we can evaluate methods]

Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42:** 795-803. [Bremer support]

Sitnikova, T., A. Rzhetsky, and M. Nei. 1995. Interior-branch and bootstrap tests of phylogenetic trees. *Molecular Biology and Evolution* **12:** 319-333. [The interior-branch test]

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. [material is in chapters 16, 19, 20, 21]

Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford. [material is in chapters 1, 2, 4, 5, section 6.4, sections 7.1 and 7.2]

Semple, C. and M. Steel. 2003. *Phylogenetics*. Oxford University Press, Oxford. [Material is in pages 204-209]