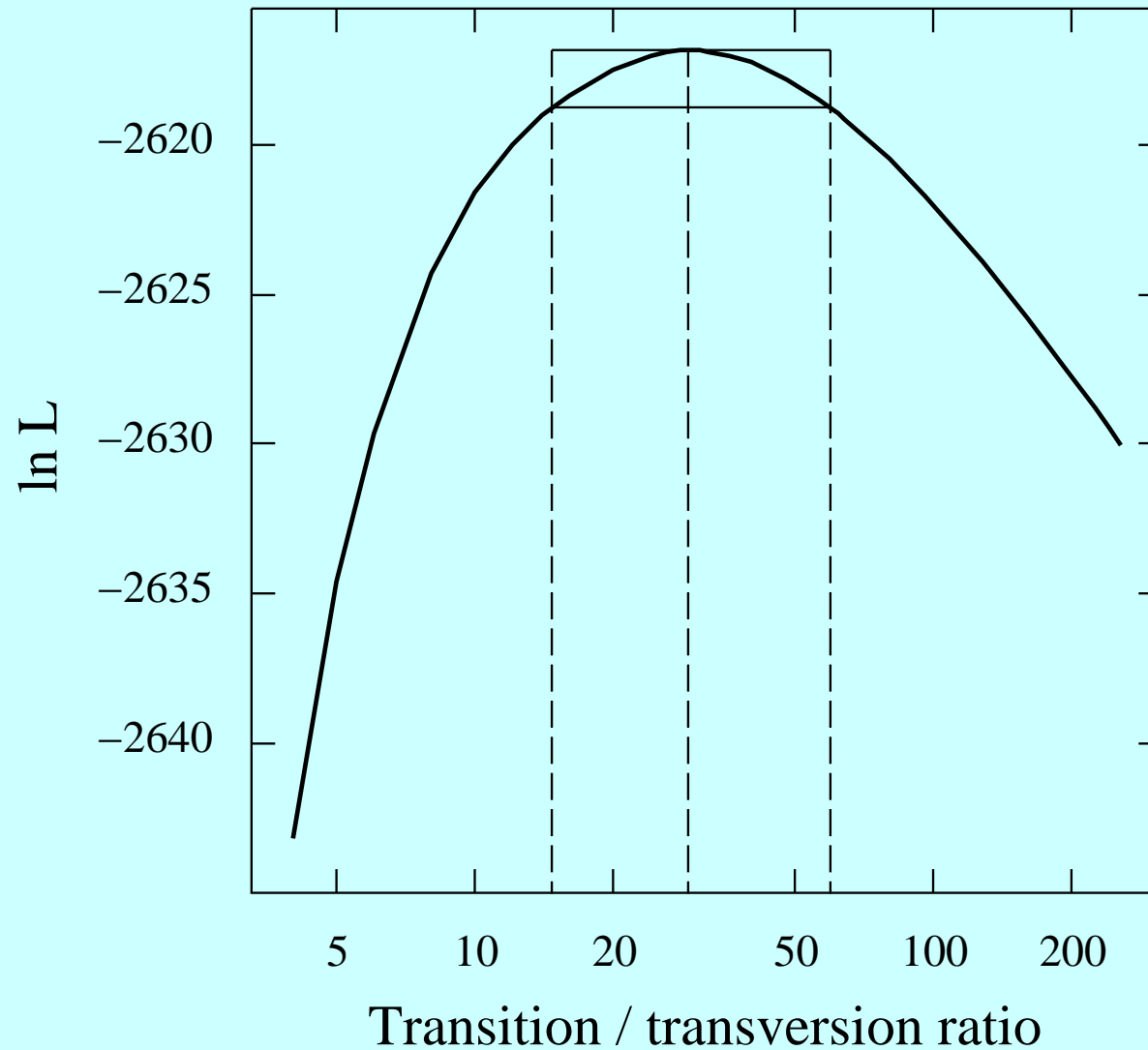# Bootstraps and testing trees

Joe Felsenstein
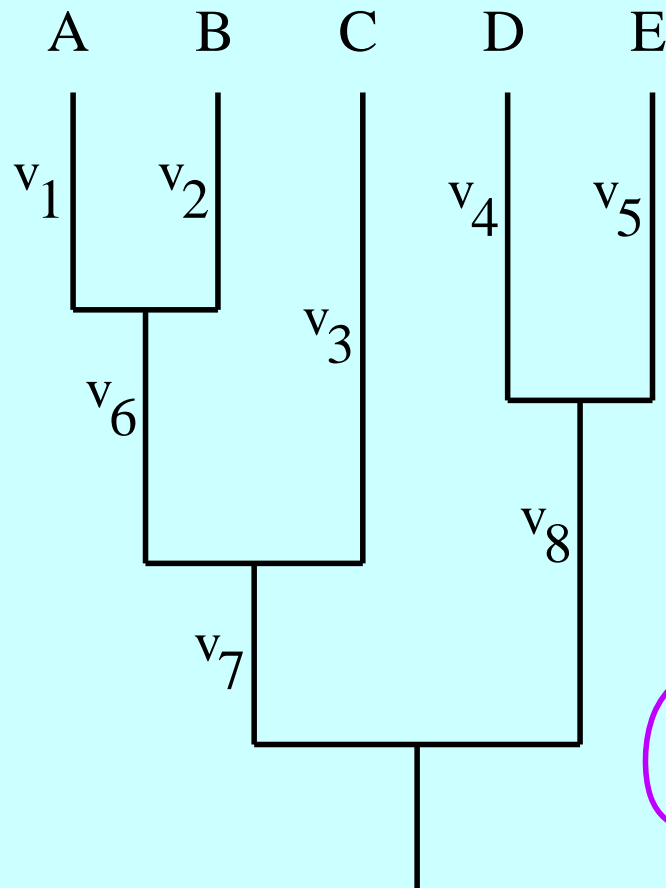
Depts. of Genome Sciences and of Biology, University of Washington

A log-likelihood curve and its confidence interval

(This is for the 14-species primates data available for download).

Constraints on a tree for a clock

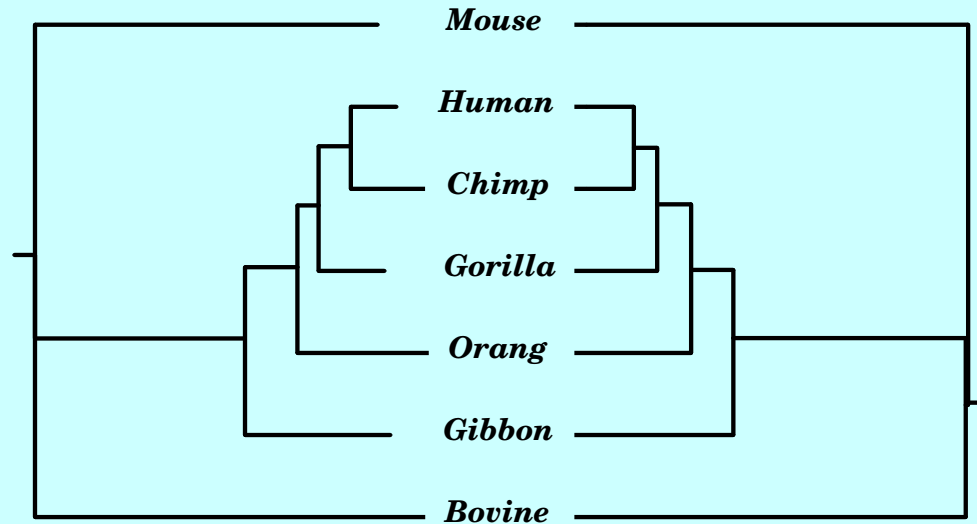A   B   C   D   E   Constraints for a clock

$$v_1 = v_2$$

$$v_4 = v_5$$

$$v_1 + v_6 = v_3$$

$$v_3 + v_7 = v_4 + v_8$$

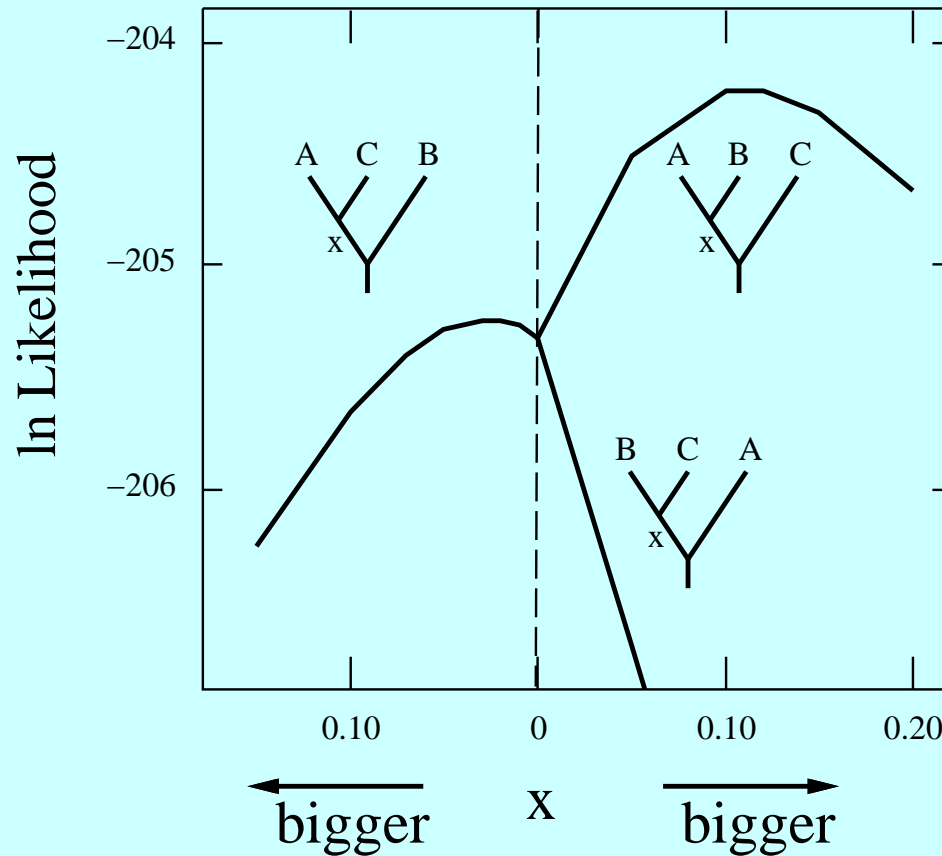Does not constrain the branch length on the unrooted tree

Likelihood-ratio test of molecular clock

	log–likelihood	parameters
Without clock	−1405.608	11
With clock	−1407.085	6
Difference	1.477	5

$\chi^2 = 2.954 \quad df = 5$

(non–significant)

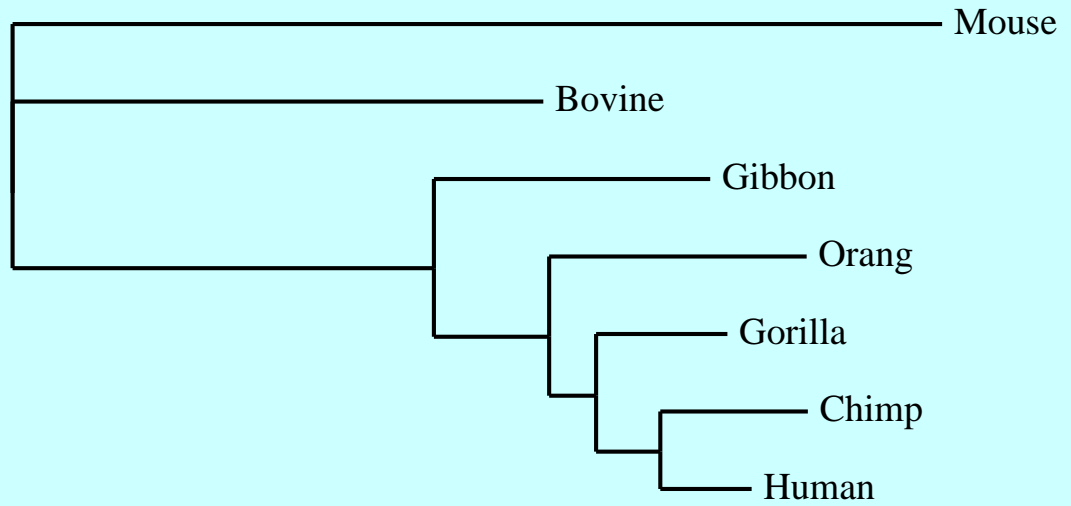(This is for this 7-species subset of the primates data).

Likelihood surface for three clocklike trees

(These are "profile likelihoods" as they show the largest likelihood for that value of x , maximizing over the other branch length in the tree.)

Two trees to be tested using KHT test

Tree I

```
                                                            ── Mouse
  ┌─────────────────────────────────────── Bovine
  │                          ┌──── Gibbon
  │                          │      ┌──── Orang
  └──────────────────────────┤      │   ┌──── Gorilla
                             └──────┤   │  ┌──── Chimp
                                    └───┤  │
                                        └──┤
                                           └──── Human
```
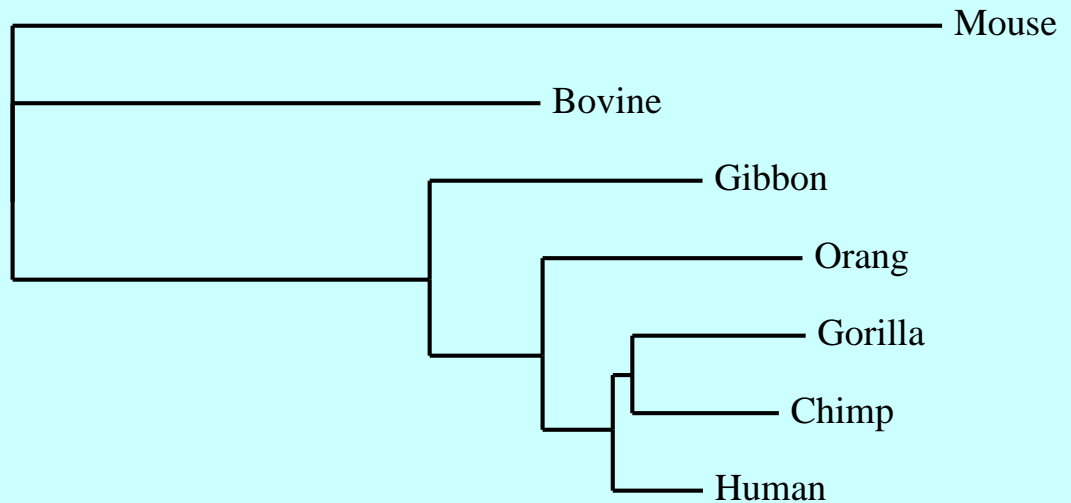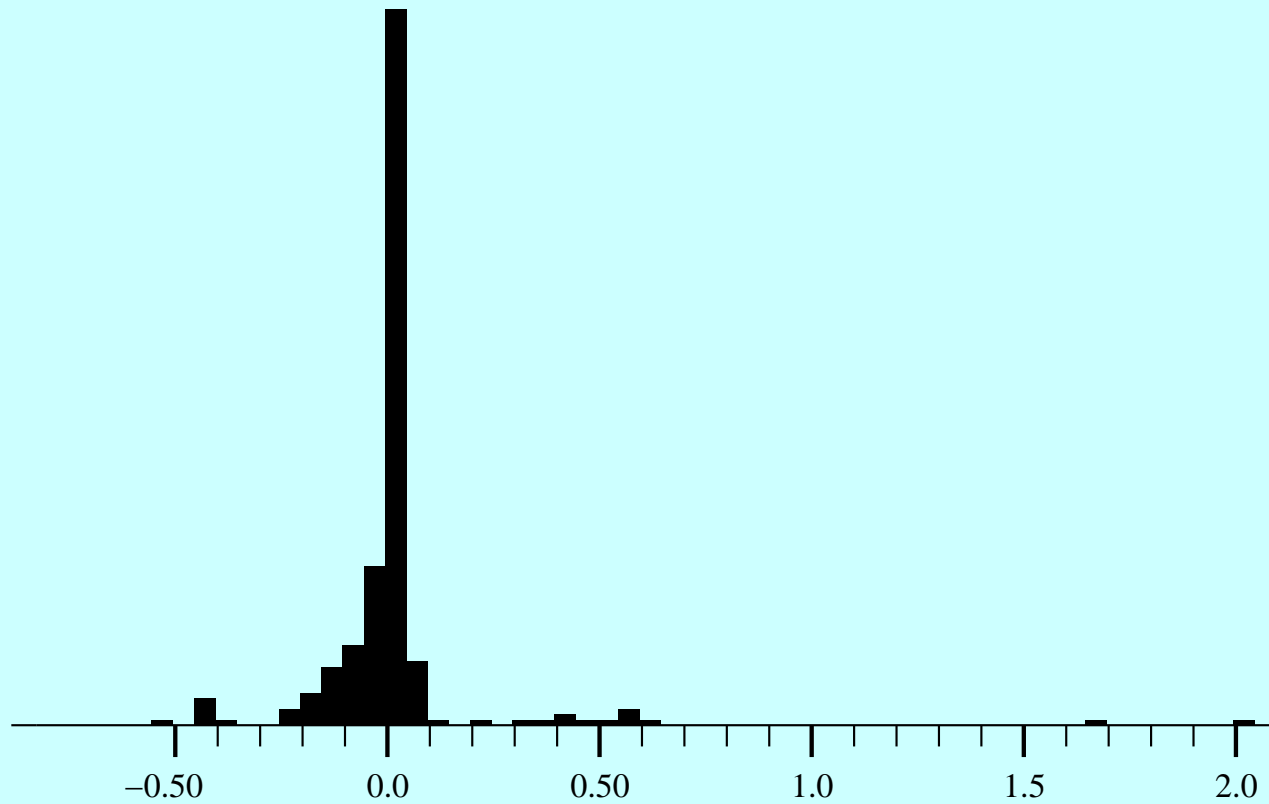
Tree II

```
                                                            ── Mouse
  ┌─────────────────────────────────── Bovine
  │                          ┌──────── Gibbon
  │                          │      ┌──────── Orang
  └──────────────────────────┤      │     ┌──── Gorilla
                             └──────┤      │
                                    └──────┤ └──── Chimp
                                           │
                                           └──── Human
```

Table of differences in log-likelihood

site\Tree	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 those differences

Difference in log likelihood at site

Do sign test, or t-test, or similar nonparametric tests.

Bootstrap sampling (with mixtures of normals)
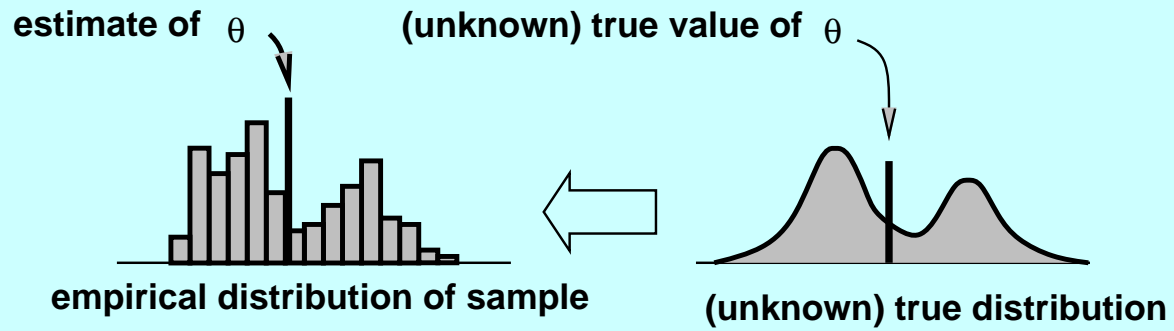
estimate of θ

(unknown) true value of θ

empirical distribution of sample

(unknown) true distribution

Bootstrap sampling (with mixtures of normals)

estimate of θ

(unknown) true value of θ

empirical distribution of sample

(unknown) true distribution

Bootstrap replicates

Distribution of estimates of parameters

Bootstrap sampling (with mixtures of normals)

estimate of θ

(unknown) true value of θ

empirical distribution of sample

(unknown) true distribution

Bootstrap replicates

Distribution of estimates of parameters

Bootstrap sampling (with mixtures of normals)

estimate of θ

(unknown) true value of θ

empirical distribution of sample

(unknown) true distribution

Bootstrap replicates

Distribution of estimates of parameters
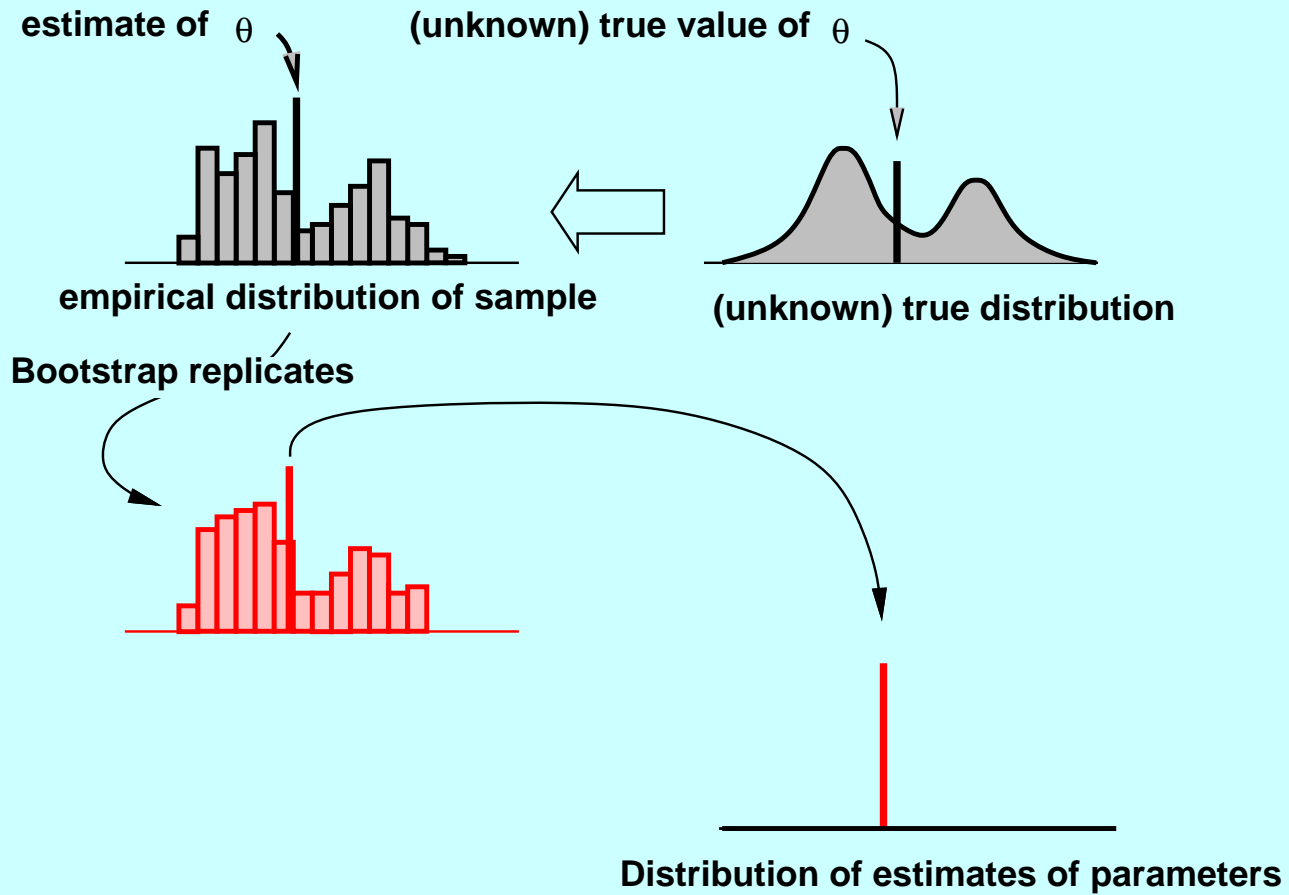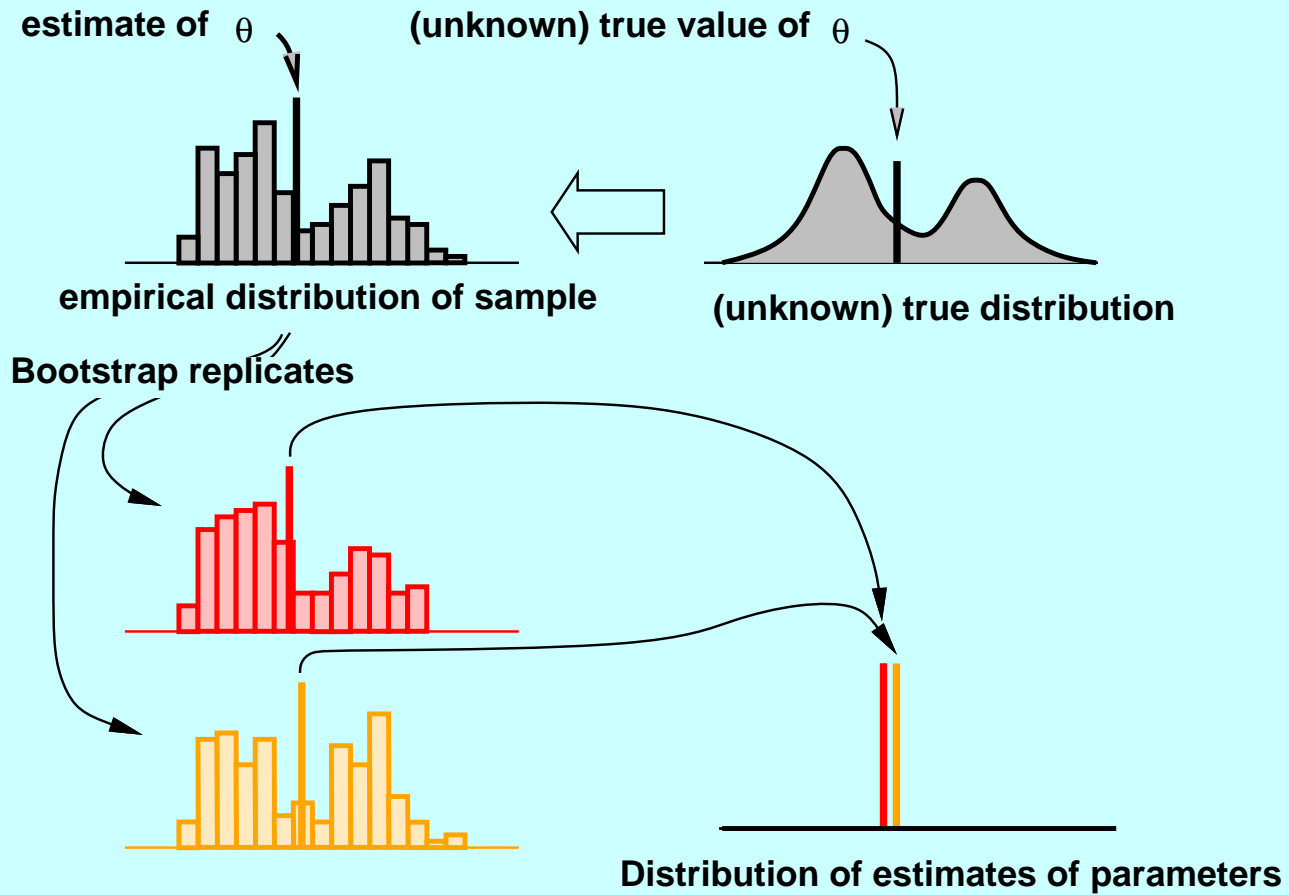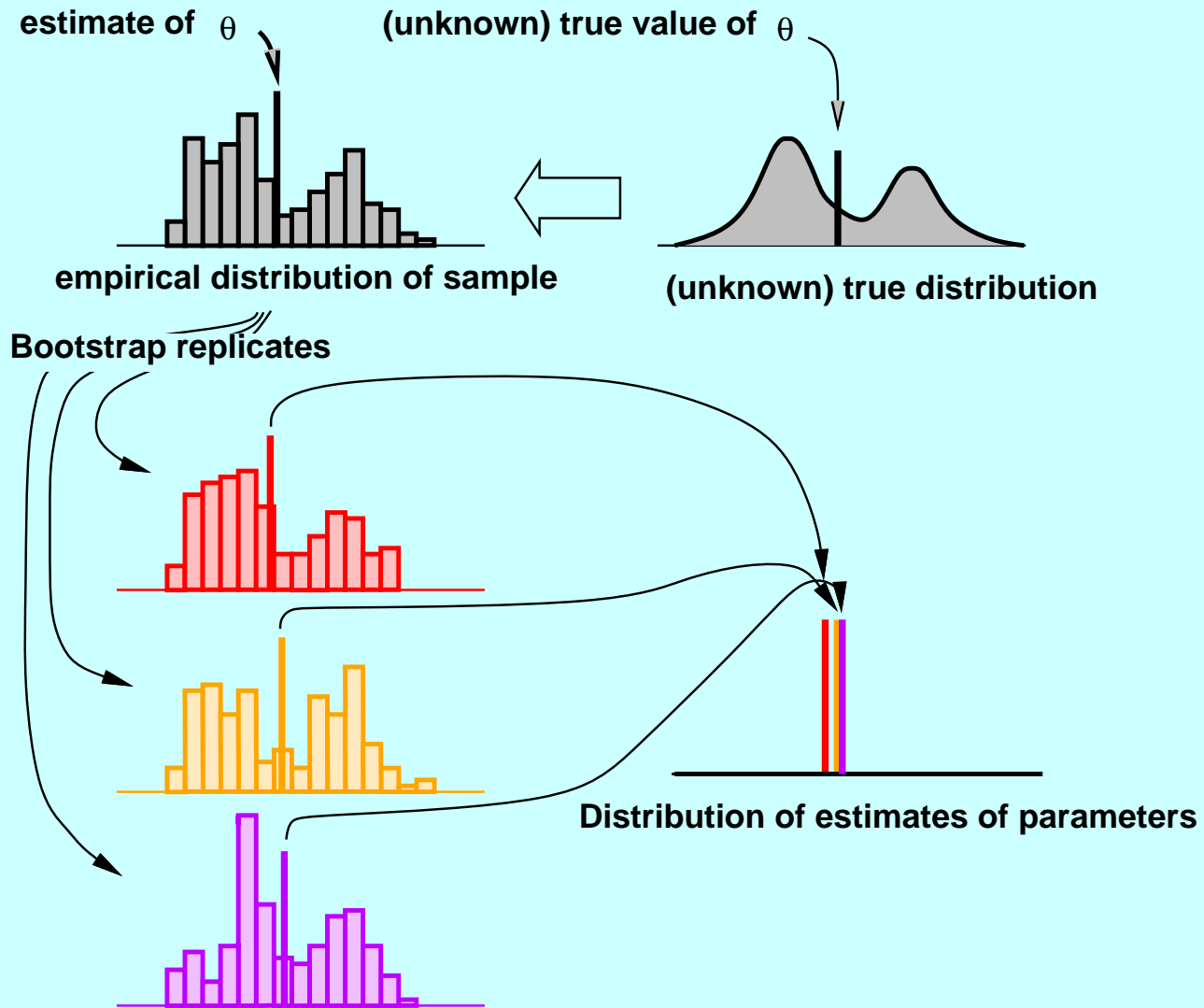
Bootstrap sampling

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)

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.

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 each of the bootstrap samples, call these $\hat{\theta}_k^*$ ($k = 1, 2, \ldots, R$)

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 each of the bootstrap samples, call these $\hat{\theta}_k^*$ ($k = 1, 2, \ldots, R$)

- 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 data points from the unknown true distribution.

Bootstrap sampling of phylogenies

Original
Data

sites

A C B D

sequences

Estimate of the tree

Bootstrap sampling of phylogenies

Original
Data

sites

A C B D

sequences

Estimate of the tree

Bootstrap
sample
#1

sites

A C B D

sequences

sample same number
of sites, with replacement

Bootstrap estimate of
the tree, #1

Bootstrap sampling of phylogenies

Original
Data

sites

sequences

Bootstrap
sample
#1

sites

sequences

sample same number
of sites, with replacement

Bootstrap
sample
#2

sites

sequences

sample same number
of sites, with replacement

(and so on)

A C B D

Estimate of the tree

A C B D

Bootstrap estimate of
the tree, #1

A C B D

Bootstrap estimate of
the tree, #2
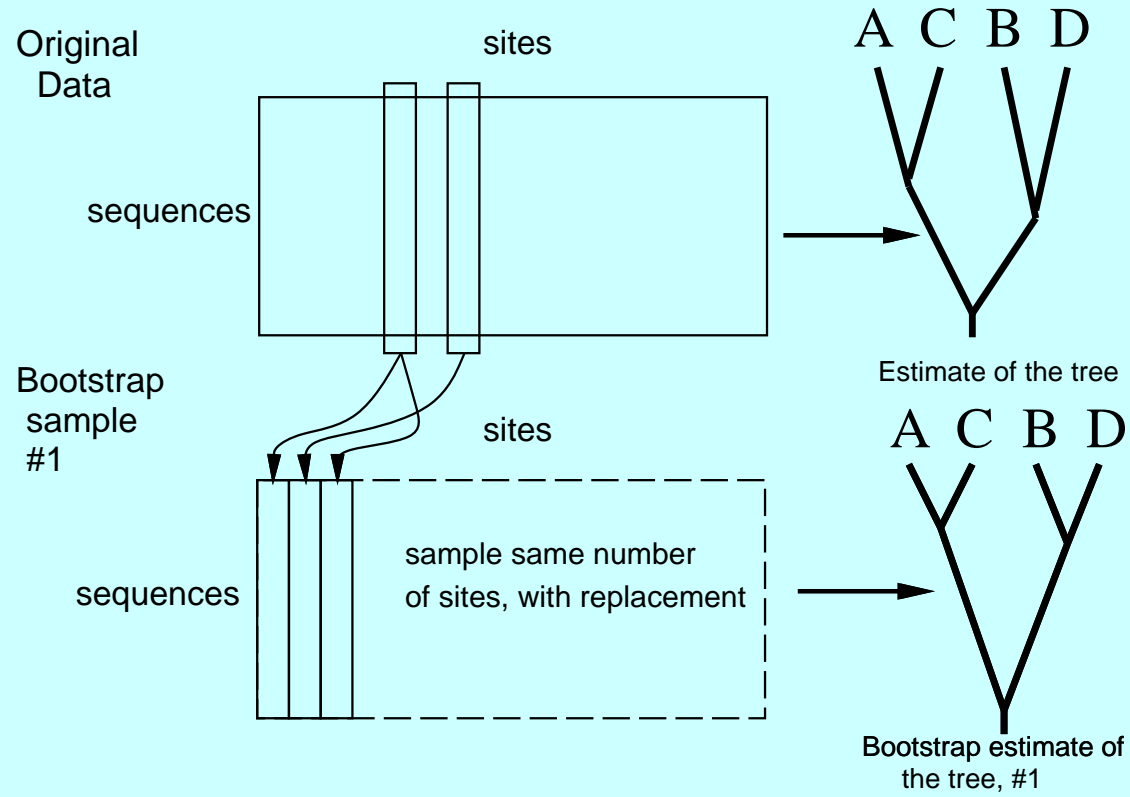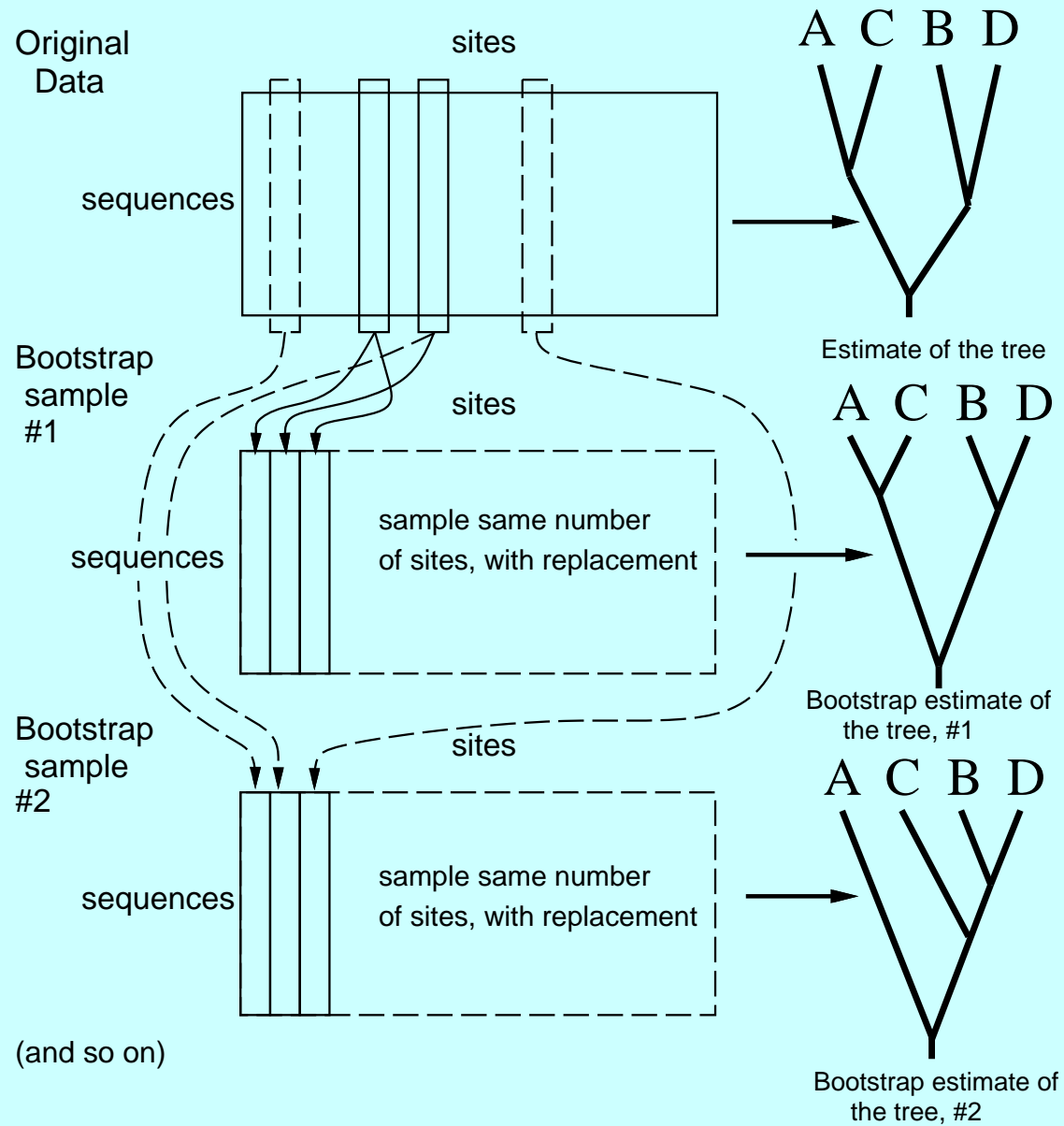
Analyzing bootstraps with phylogenies

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.

There are many possible ways. The one I will describe here is the most useful, but not the only way we could go.

Analyzing bootstraps with 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.

There are many possible ways. The one I will describe here is the most useful, but not the only way we could go.

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.

Partitions from branches in an (unrooted) tree
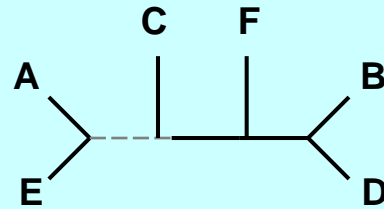
AE | BCDF

Partitions from branches in an (unrooted) tree
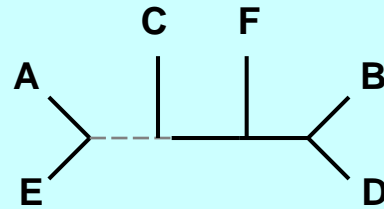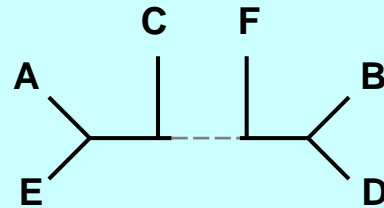
AE | BCDF

ACE | BDF

Partitions from branches in an (unrooted) tree
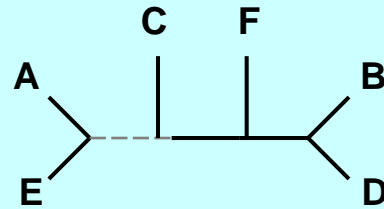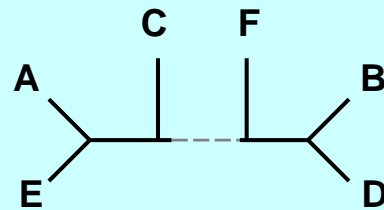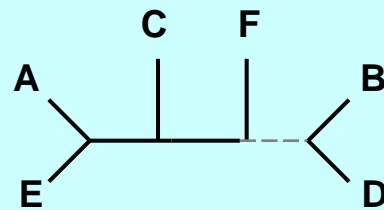
AE | BCDF

ACE | BDF

ACEF | BD

Partitions from branches in an (unrooted) tree
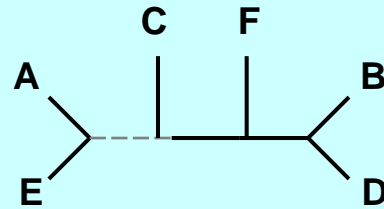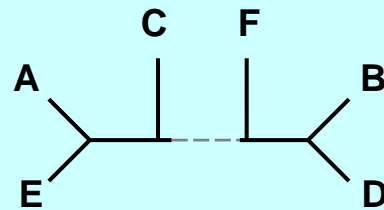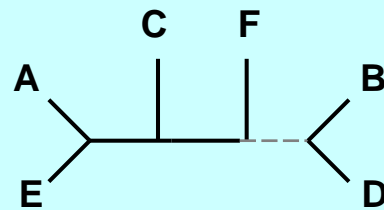
AE | BCDF
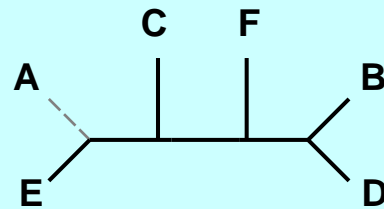
ACE | BDF

ACEF | BD

A | CEFBD

Partitions from branches in an (unrooted) tree
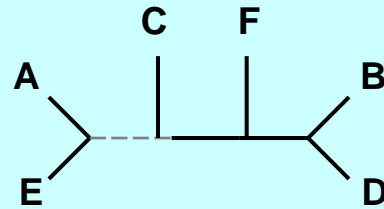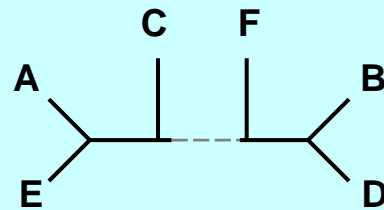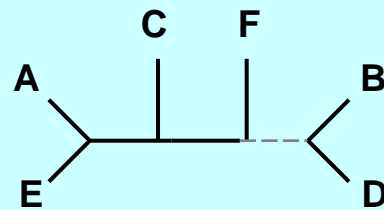
AE | BCDF
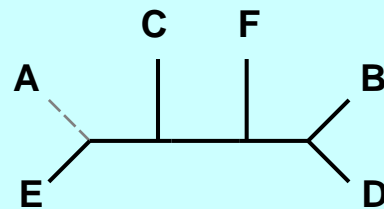
ACE | BDF

ACEF | BD

A | CEFBD

and so on for all the other external (tip) branches

The majority-rule consensus tree
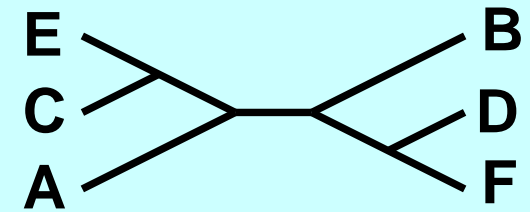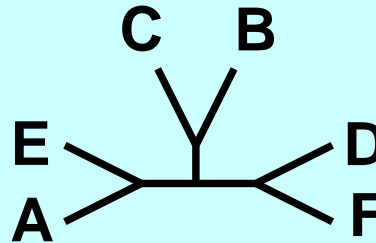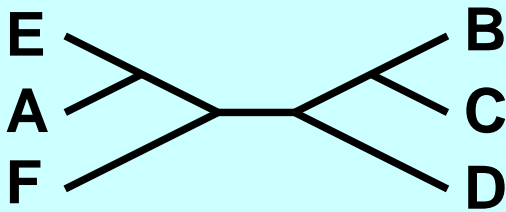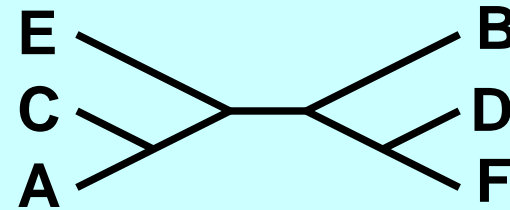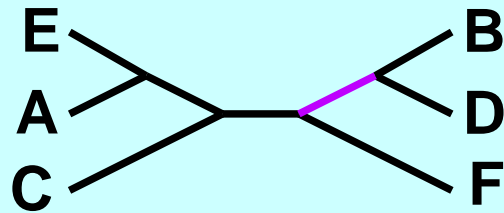
Trees:

How many times each (non–tip) partition of species is found:

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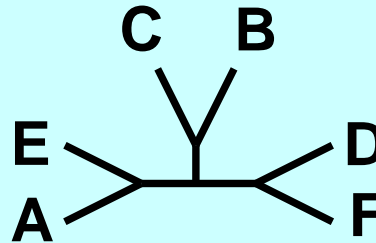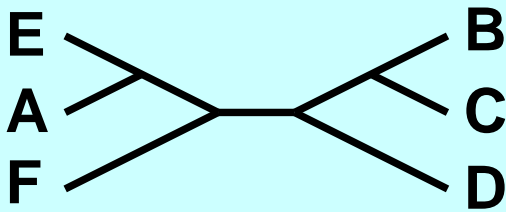**ABCE | DF**

The majority-rule consensus tree
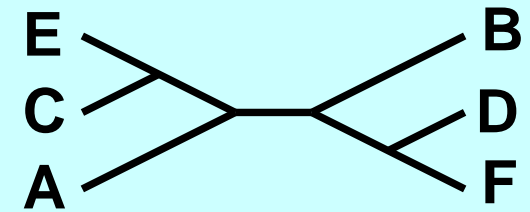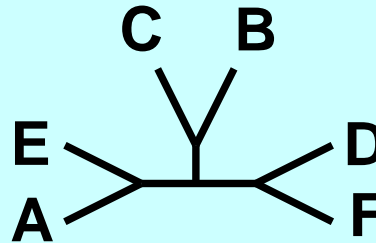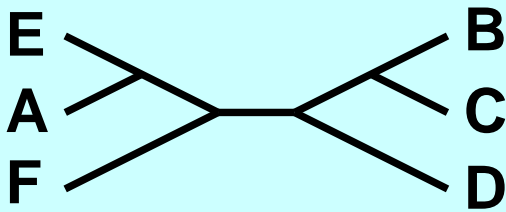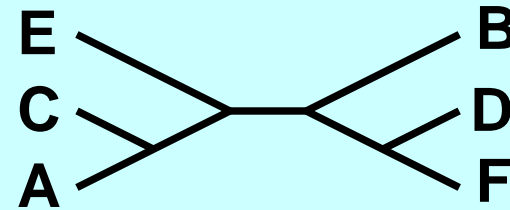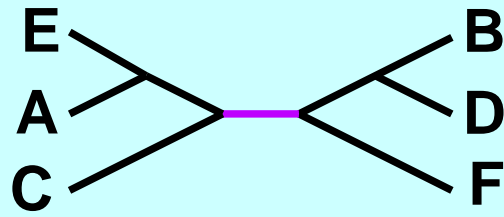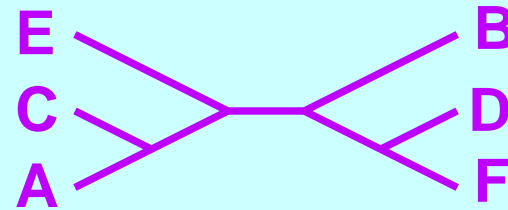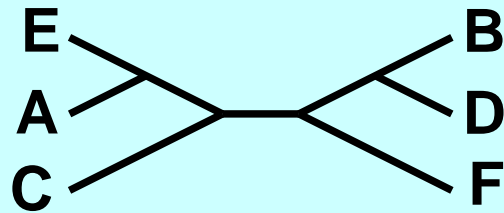
Trees:

How many times each (non–tip) partition of species is found:

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

The majority-rule consensus tree
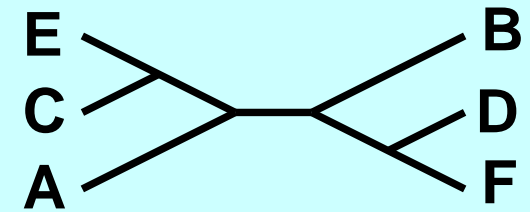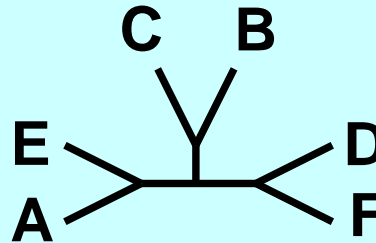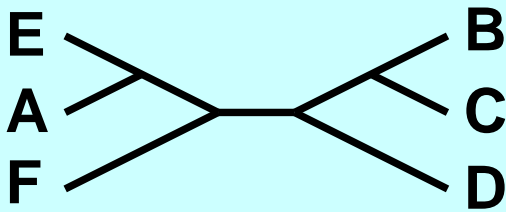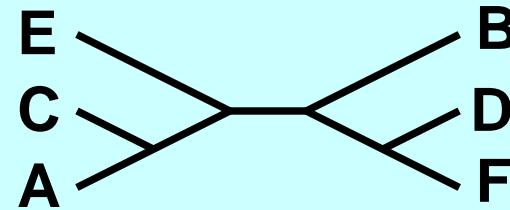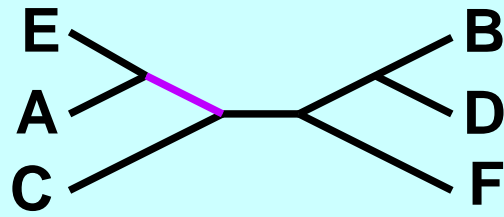
Trees:

How many times each (non–tip) 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 majority-rule consensus tree
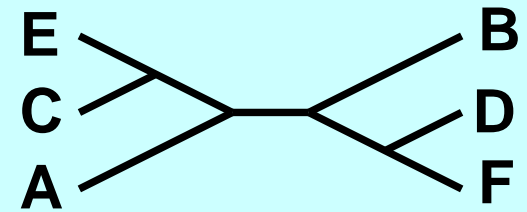
Trees:

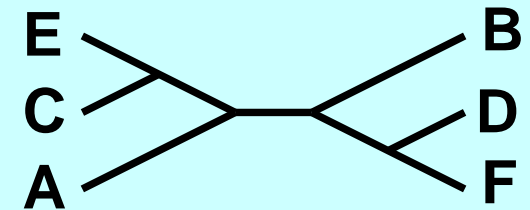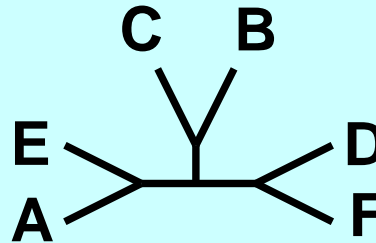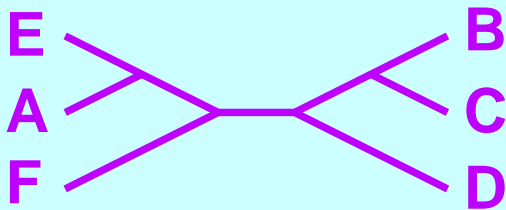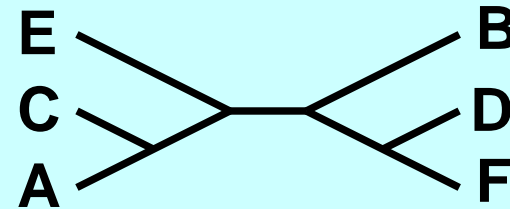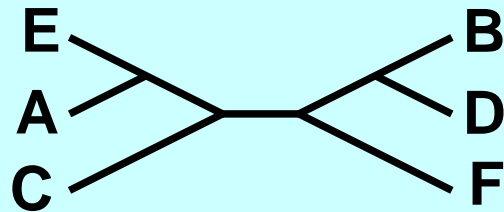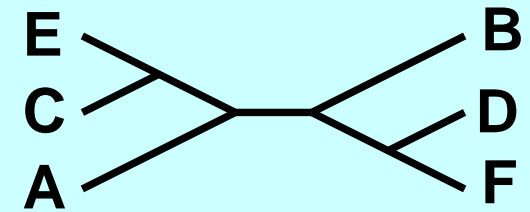How many times each (non–tip) partition of species is found:

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

The majority-rule consensus tree

Trees:

How many times each (non–tip) 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 majority-rule consensus tree
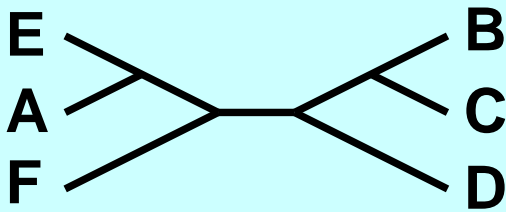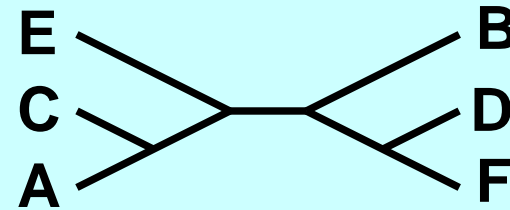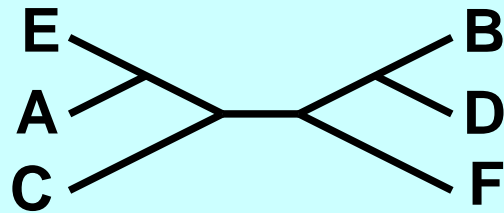
Trees:

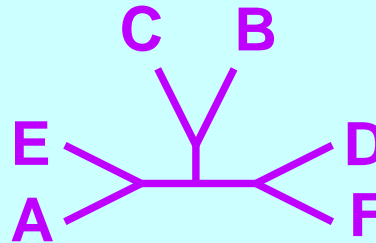How many times each (non–tip) partition of species is found:

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

The majority-rule consensus tree

Trees:

How many times each (non–tip) 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 majority-rule consensus tree
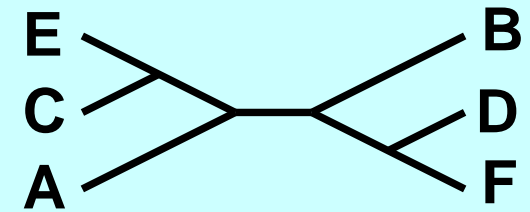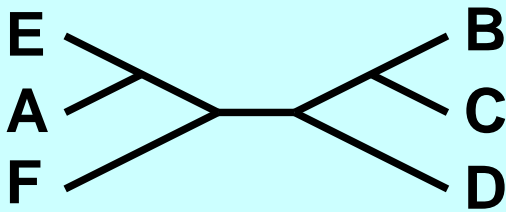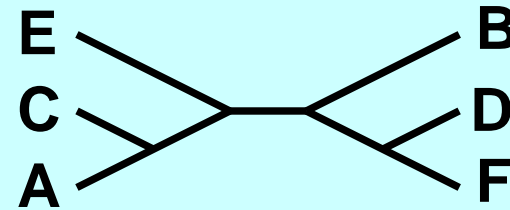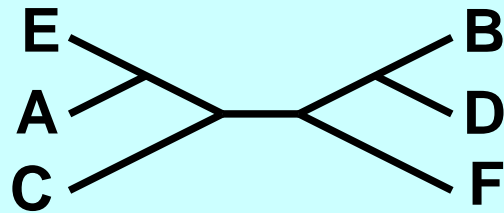
Trees:
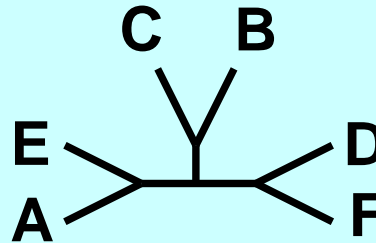

How many times each (non–tip) partition of species is found:

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 majority-rule consensus tree

Trees:

How many times each (non–tip) partition of species is found:

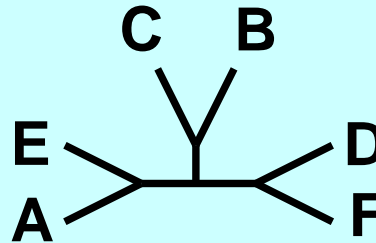|	
AE | BCDF	**3**
ACE | BDF	**3**
ACEF | BD	1
AC | BDEF	1
AEF | BCD	1
ADEF | BC	2
ABDF | EC	1
ABCE | DF	**3**

Majority–rule consensus tree of the unrooted trees:
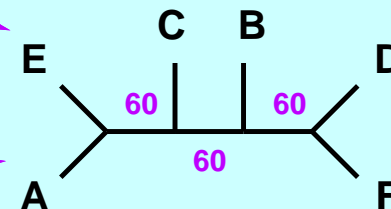
Bootstrap sampling of a phylogeny
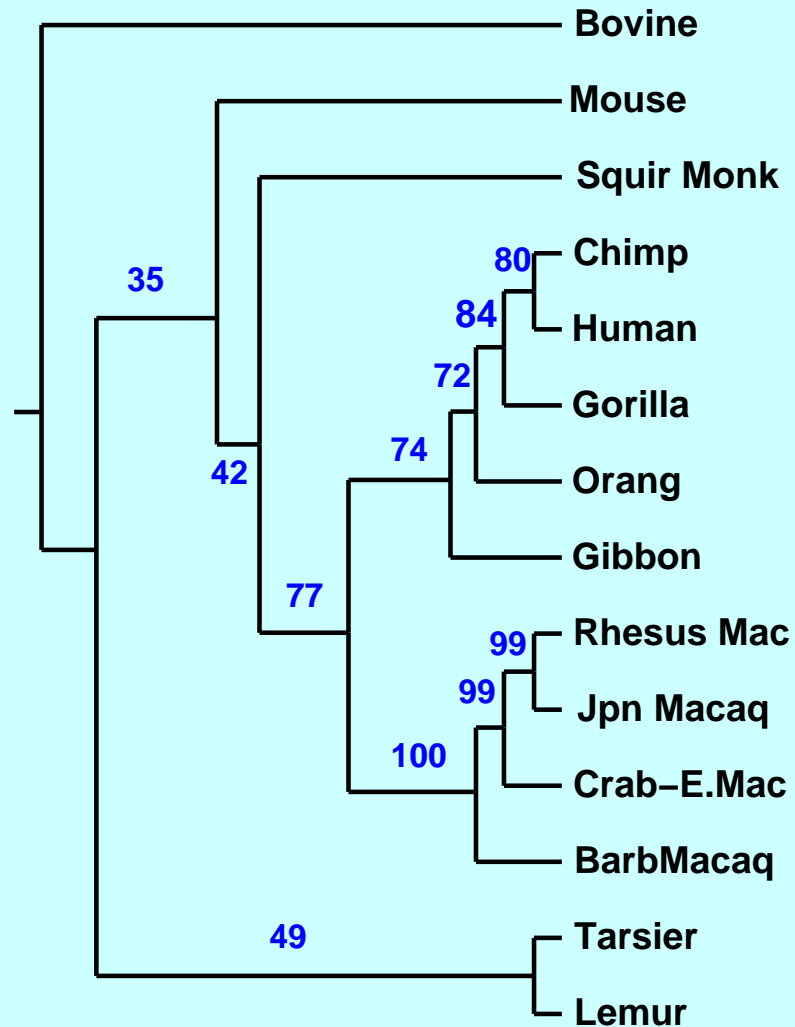
In this example, parsimony was used to infer the tree.

Potential problems with the bootstrap

- Sites may not evolve independently

Potential problems with the bootstrap

- Sites may not evolve independently

- Sites may not come from a common distribution (but you can consider them to be sampled from a mixture of possible distributions)

Potential problems with the bootstrap

- Sites may not evolve independently

- Sites may not come from a common distribution (but you can consider them to be sampled from a mixture of possible distributions)

- If do not know which branch is of interest at the outset, a "multiple-tests" problem means that the most extreme P values are overstated

Potential problems with the bootstrap

- Sites may not evolve independently

- Sites may not come from a common distribution (but you can consider them to be sampled from a mixture of possible distributions)

- If do not know which branch is of interest at the outset, a "multiple-tests" problem means that the most extreme P values are overstated

- P values are biased (too conservative)

Potential problems with the bootstrap

- Sites may not evolve independently

- Sites may not come from a common distribution (but you can consider them to be sampled from a mixture of possible distributions)

- If do not know which branch is of interest at the outset, a "multiple-tests" problem means that the most extreme P values are overstated
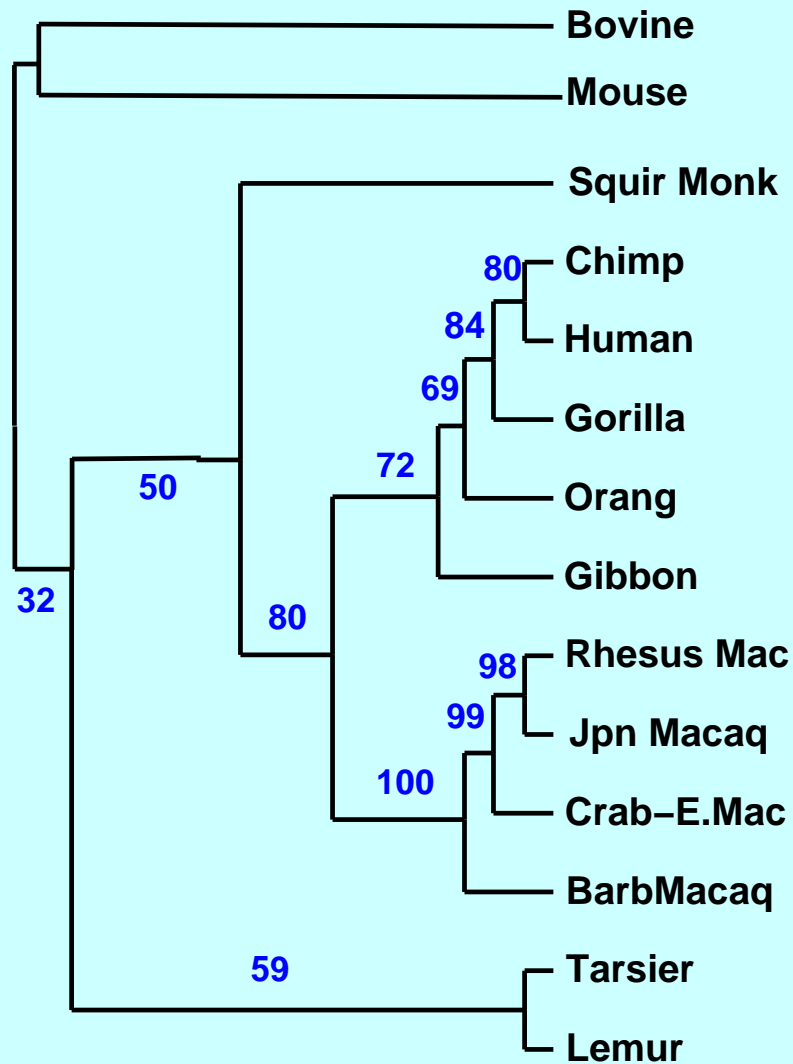
- P values are biased (too conservative)

- Bootstrapping does not correct biases in phylogeny methods
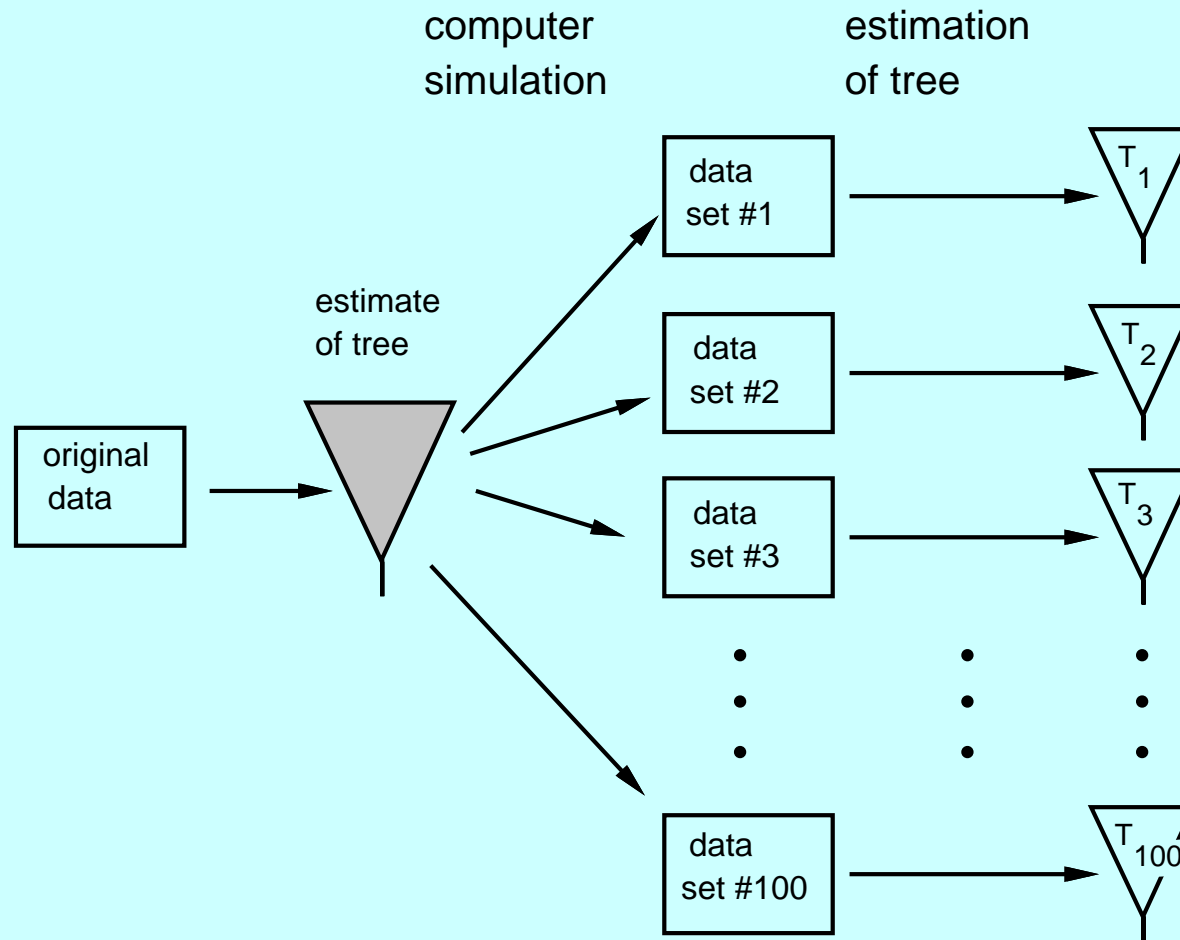
Delete-half jackknife P values

In this example, parsimony was used to infer the tree.

A diagram of the parametric bootstrap

References

Bootstraps etc.

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]**

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]**

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

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

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

(more references)

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:** 221-224. **[The first paper on the KHT test]**

Goldman, N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36:** 182-98. **[Parametric bootstrapping for testing models]**

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

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]**

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]**