

# Statistical methods for assessing confidence in phylogenies

by Andres Figueroa

Computer Science Department, UCR. 1

## Building phylogenies

- The purpose of a phylogeny is to illustrate how a group of objects (usually genes or organisms) are related to one another

2

## Kinds of phylogenies

- Cladogram**
  - shows relative recency of common ancestry.
- Phylograms**
  - depict the amount of evolutionary change that has occurred along the different branches.
- Dendograms**
  - depict the times of divergence.

3

## Types of phylogenies

unrooted

rooted

Only specifies relationships not the evolutionary path Root R is a common ancestor of all species

4

## Types of data

sequences

	sites						
	1	2	3	4	5	6	7
1	T	T	A	T	T	A	A
2	A	A	T	T	T	A	A
3	A	A	A	A	A	T	A
4	A	A	A	A	A	A	T

distances

2	3
3	5 4
4	5 4 2
	1 2 3

5

## Classes of algorithms used to infer phylogenies

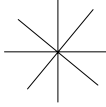
- Distance methods.**
  - UPGMA (Sneath et al. 1973).

6

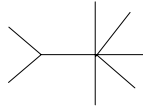
## Classes of ... (cont.)

- NJ (Saitou et al. 1987). Identify neighbors that sequentially minimize the total length of the tree.

1. Start with a star tree – no topology



2. Choose a pair of objects that minimize total branch lengths in the tree.



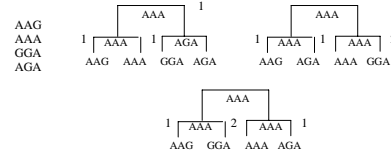
- This pair is reconsidered as single object and distance matrix is recalculated.
- Next pair of objects that gives smallest branch length is chosen.
- Iterate until complete.

7

## Classes of ... (cont.)

- Maximum Parsimony (Camin et al. 1965).

- The best tree is the one which requires the least number of substitutions.



8

## Classes of ... (cont.)

- Number of bifurcating unrooted trees and rooted trees for  $n$  objects. (Felsenstein 1978)

$$\prod_{k=3}^n (2k-5) = \frac{(2n-5)!}{2^{n-3}(n-3)!}; n \geq 3$$

$$\prod_{k=2}^n (2k-3) = \frac{(2n-3)!}{2^{n-2}(n-2)!}; n \geq 2$$

9

## Classes of ... (cont.)

- Maximum Likelihood (Felsenstein 1981)

- Given some data  $D$ , and hypothesis  $H$ , the likelihood of the data is given by  $L_D = \Pr(D|H)$  which is the probability of obtaining  $D$  given  $H$ .

$D$  = data

$H$  = tree

$P(b|a,t)$  denote the probability of a residue  $a$  having being substituted by a residue  $b$  over the branch  $t$ .

10

How confident am I that my tree is correct?

11

Most scientific measures are accompanied by some estimate of precision

- For example  $30 \pm 0.3$  cm.
- Phylogenies should also be accompanied by some indication of confidence limits.
- One reason for a poor estimate is sampling error due to the data.
- As a consequence, estimates of phylogeny based on samples will be accompanied by error.

12

## Assessing confidence in phylogenies

- Several methods have been proposed that attach numerical values to internal branches in trees that are intended to provide some measure of the strength of support for those branches and the corresponding groups.
- These methods include:
  - character resampling methods - the bootstrap and jackknife.

13

## Bootstrap (Efron 1979)

- Bootstrapping is a modern statistical technique that uses computer intensive random resampling of data to determine sampling error or confidence intervals for some estimated parameter.

14

## Estimating the sampling error

- A good way to measure sampling error is to take multiple samples from the population being studied and compare the estimates from the different samples.
- The spread of the estimates gives an indicator, *i.e.* how much our conclusions would vary depending on the samples we took.

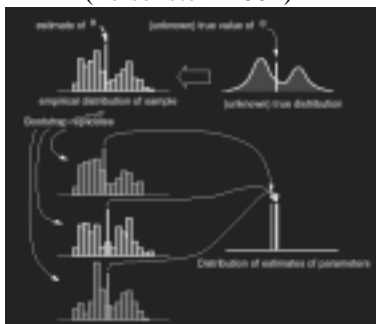
15

## Sampling error using the bootstrap

- The bootstrap invokes the same underlying principle, but rather than resample replicates from the population we resample pseudoreplicates from the data.
- For each pseudoreplicate we derive an estimate of the parameter we are trying to measure. The variation among the estimates derived from each pseudoreplicate provides a measure of the sampling error.

16

## Sampling error using the bootstrap (Felsenstein 2002)



17

## The bootstrap algorithm for estimating standard errors

To infer an error in a quantity,  $\theta$ , estimated by  $\hat{\theta}$  from a sample  $\mathbf{x} = \{x_1, \dots, x_n\}$

- Select  $B$  independent bootstrap samples  $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \dots, \mathbf{x}^{*B}$ , each consisting of  $n$  data values drawn with replacement from  $\mathbf{x}$ .
- Evaluate the bootstrap replication corresponding to each bootstrap sample,

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}); b = 1, 2, \dots, B.$$

- Estimate the standard error  $se_{\hat{\theta}}$  by the sample standard deviation of the  $B$  replications<sup>ns</sup>

$$se_{\hat{\theta}} = \left\{ \sum_{b=1}^B \left[ \hat{\theta}^*(b) - \hat{\theta}(\cdot) \right]^2 / (B-1) \right\}^{1/2}$$

where  $\hat{\theta}(\cdot) = \sum_{b=1}^B \hat{\theta}^*(b) / B$

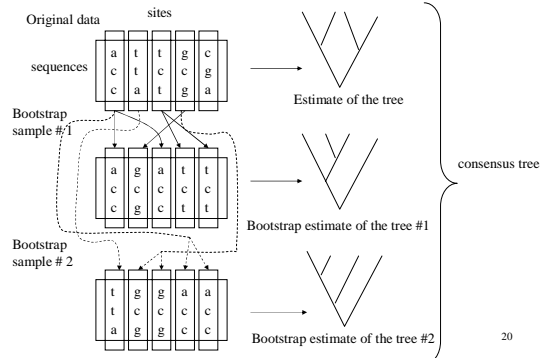
18

## Bootstrapping phylogenies (Felsenstein 1985)

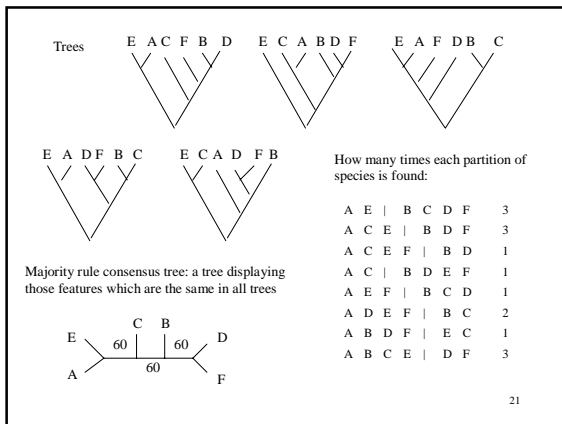
- The input is  $n$  sequences.
- Sites are resampled with replacement to create many bootstrap replicate data sets.
- Each bootstrap replicate data set is analysed (e.g. with parsimony, distance, ML etc.)
- Agreement among the resulting trees is summarized with a consensus tree.
- Frequencies of occurrence of groups, bootstrap proportions (BPs), are a measure of support for those groups.

19

## Bootstrapping phylogenies (cont.)



20



21

## Bootstrap - interpretation

- Bootstrapping is a very valuable and widely used technique.
- The interpretation of bootstrap proportions (BPs) depends on the assumption that the original data is a random sample from a much larger set of independent and identically distributed data.
- BPs give an idea of how likely a given branch would be unaffected if additional data, with the same distribution, became available.

22

## Bootstrap – interpretation (cont.)

- BPs are not the same as confidence intervals. There is no agreement about what constitutes a “good” bootstrap value (> 70%, > 80%, > 85% ?).
- If the estimated tree is inconsistent any bootstraps won’t help you.
- Low BPs doesn’t mean the relationship is false, only that it is poorly supported.

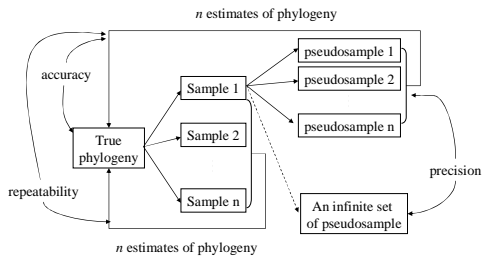
23

## Accuracy, Repeatability, and Precision (D. Hillis et al. 1993)

- Before evaluating the performance of bootstrapping, it is necessary to distinguish among the terms “repeatability”, “accuracy”, and “precision.”
- Repeatability
  - The probability that a specified group will be found in an analysis of an independent sample of characters.
- Accuracy
  - The probability that a specified group is contained in the true phylogeny.
- Precision
  - The degree to which BPs based on a finite set of pseudosamples are expected to match the values that would be obtained from an infinite set of pseudosamples.

24

### Accuracy, Repeatability, and Precision (cont.)



25

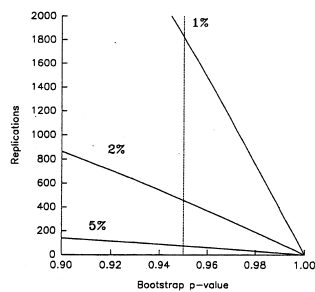
### How many replications is enough? (S. Blair Hedges 1992)

- BP is the proportion of trees containing a particular phylogenetic group.
- BP follows a binomial distribution (P,n), which has a variance of

$$\sigma^2 = P(1-P)/n$$

26

### How many replications is enough?



27

### Jackknife (Miller 1974)

- Jackknifing is very similar to bootstrapping and differs only in the character resampling strategy
- Some proportion of sites are randomly deleted (e.g. 50% known as delete-half jackknife.)
- Jackknifing is not as widely available or widely used as bootstrapping
- Tends to produce broadly similar results

28

Thanks

Questions?

29

### References

- J. Camin and R. Sokal, *A method for deducing branching sequences in phylogeny*, *Evolution*, 19:311-326, 1965.
- P. Sneath, and R. Sokal, *Numerical Taxonomy*, pages 230-234. W.K. Freeman and Company, San Francisco, CA, 1973.
- R. Miller, *The Jackknife - a review*, *Biometrika*, 61:1-15, 1974.
- J. Felsenstein, *The number of evolutionary trees*, *Systematic Zoology*, 27:23-33, 1978.
- B. Efron, *Bootstrap methods: another look at the jackknife*, *Annals of Statistics*, 7:1-26, 1979.
- J. Felsenstein, *Evolutionary trees from DNA sequences: a maximum likelihood approach*, *J. Mol. Evol.*, 17: 368-376, 1981.
- J. Felsenstein, *Confidence limits on phylogenies: an approach using the bootstrap*, *Evolution*, 39(4):783-791, 1985.
- N. Saitou, and M. Nei, *The Neighbor-joining method: A new method for reconstructing phylogenetic trees*, *Mol. Biol. Evol.*, 4:406-425, 1987.
- S. Blair Hedges, *The number of replications needed for accurate estimation of the bootstrap P value in phylogenetic studies*, *Mol. Biol. Evol.*, 9(2):366-369, 1992.
- D. Hillis and J. Bull, *An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis*, *Systematic Biology*, 42(2):182-192, 1993.

30