

Bootstrap Methods in Phylogenetics

Emmanuel Paradis

Institut Pertanian Bogor

December 12, 2011

Uncertainty in data analyses is ***crucial***.

Simple examples:

1. Flipping a coin 10 times: 4 heads; 10 other times: 7 heads
Estimates: $\hat{p} = 0.4$; $\hat{p} = 0.7$

Uncertainty in data analyses is ***crucial***.

Simple examples:

1. Flipping a coin 10 times: 4 heads; 10 other times: 7 heads

Estimates: $\hat{p} = 0.4$; $\hat{p} = 0.7$

2. Simulating random normal variates with R:

```
> mean(rnorm(100))
```

```
[1] 0.01393119
```

```
> mean(rnorm(100))
```

```
[1] 0.1267855
```

```
> mean(rnorm(100))
```

```
[1] -0.09425586
```

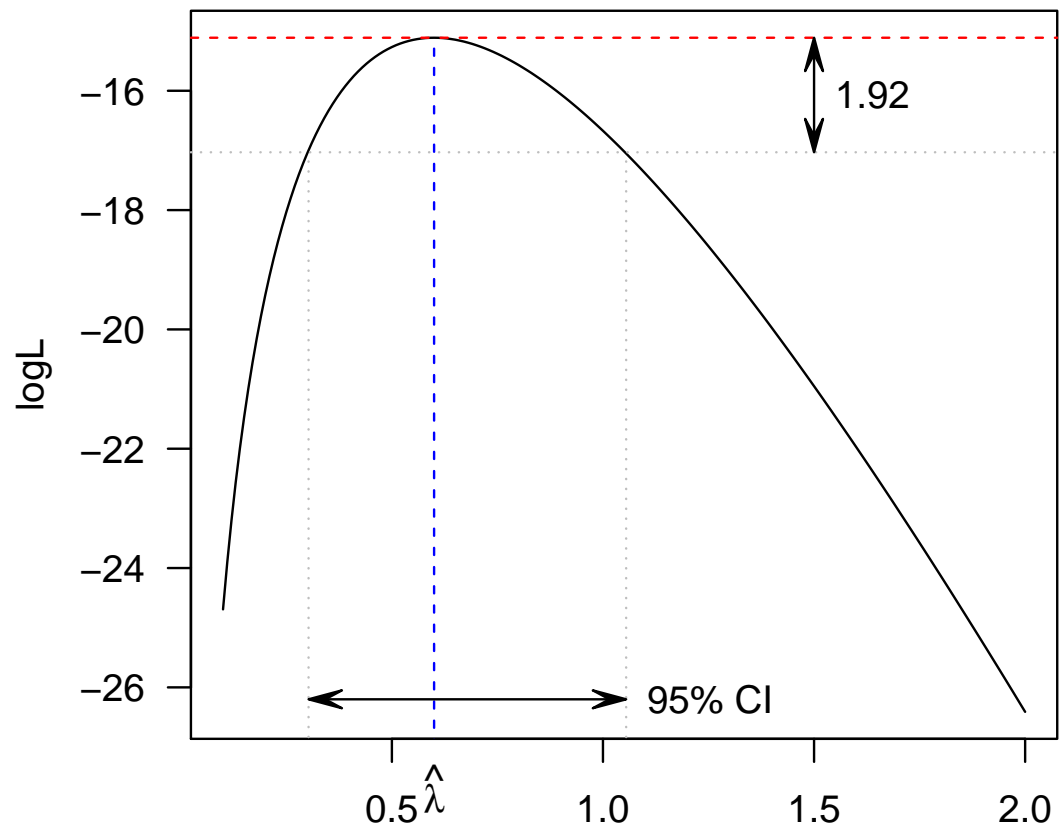
Some more vital examples:

1. Assessing the effects of a new medicine from a limited number of clinical trials.
2. Assessing the effects of mining, logging, pesticides, etc, on natural populations of animals and plants.
3. Predicting the outcome of an epidemic (H1N1 in Europe in 2009).

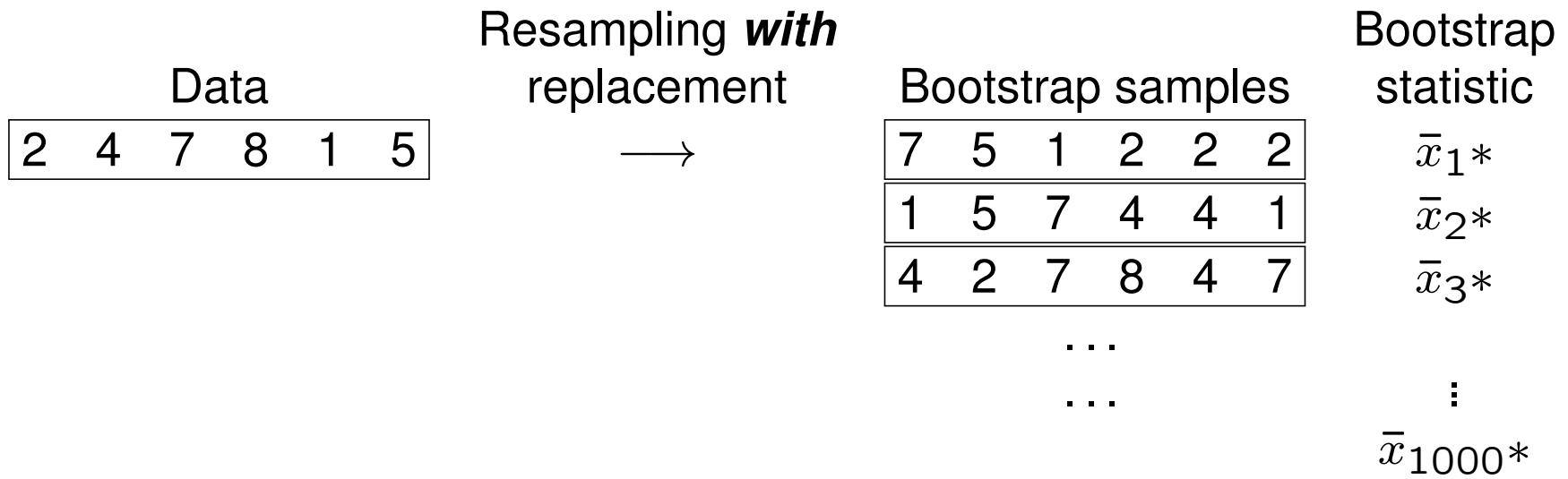
Two theories of uncertainties:

- ▶ The Theorem of the Central Limit (TCL): the estimator $\hat{\theta}$ of a parameter θ follows a normal distribution.
Example of the flipping coin: $\hat{p} \sim \mathcal{N}(0.5, \sigma_n^2)$
- ▶ Maximum Likelihood (ML): more general.

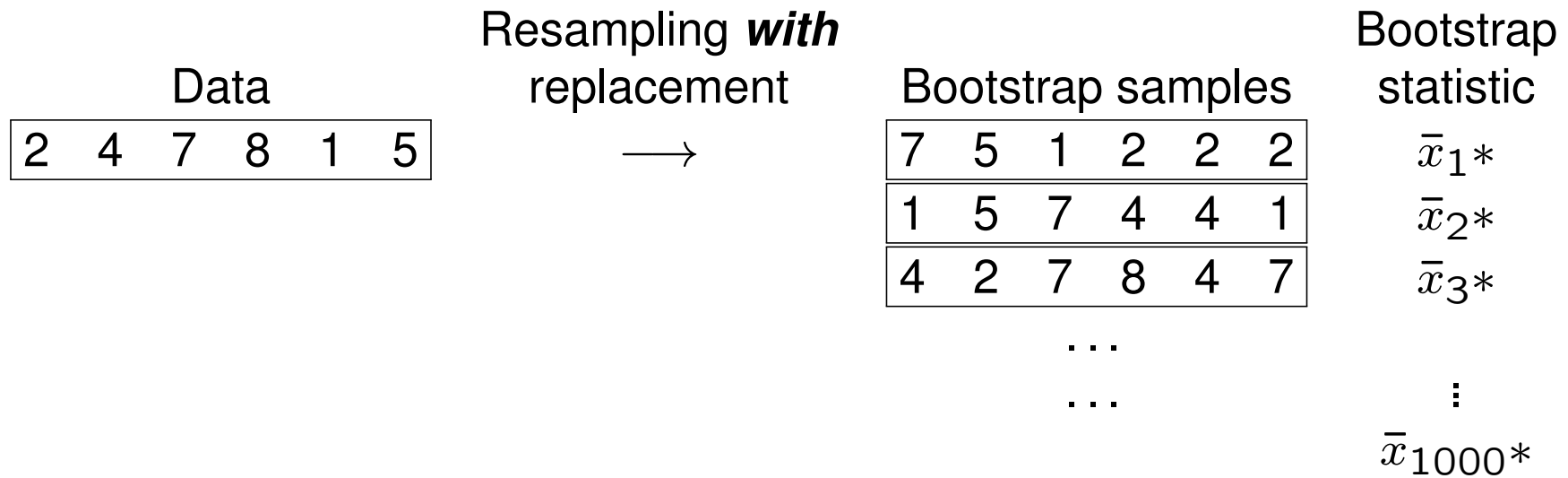
Confidence interval with profile likelihood



Principle of the ***bootstrap***:



Principle of the **bootstrap**:



$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{2+4+7+8+1+5}{6} = 4.5$$

$$SE_{\text{BOOT}}(\hat{\mu}) = \sqrt{\text{var}(\bar{x}^*)}$$

“Manual” bootstrap:

```
> sd(replicate(99999, mean(sample(x, replace = TRUE))))  
[1] 1.020867
```

“Manual” bootstrap:

```
> sd(replicate(99999, mean(sample(x, replace = TRUE))))  
[1] 1.020867
```

With the package boot in R:

```
> library(boot)  
> f <- function(x, w) mean(x * w)  
> boot(x, f, R = 99999, stype = "f")
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

```
boot(data = x, statistic = f, R = 99999, stype = "f")
```

Bootstrap Statistics :

| | original | bias | std. error |
|-----|----------|--------------|------------|
| t1* | 4.5 | 0.0006200062 | 1.021678 |

Call:

```
boot(data = x, statistic = f, R = 99999, stype = "f")
```

Bootstrap Statistics :

| | original | bias | std. error |
|-----|----------|--------------|------------|
| t1* | 4.5 | 0.0006200062 | 1.021678 |

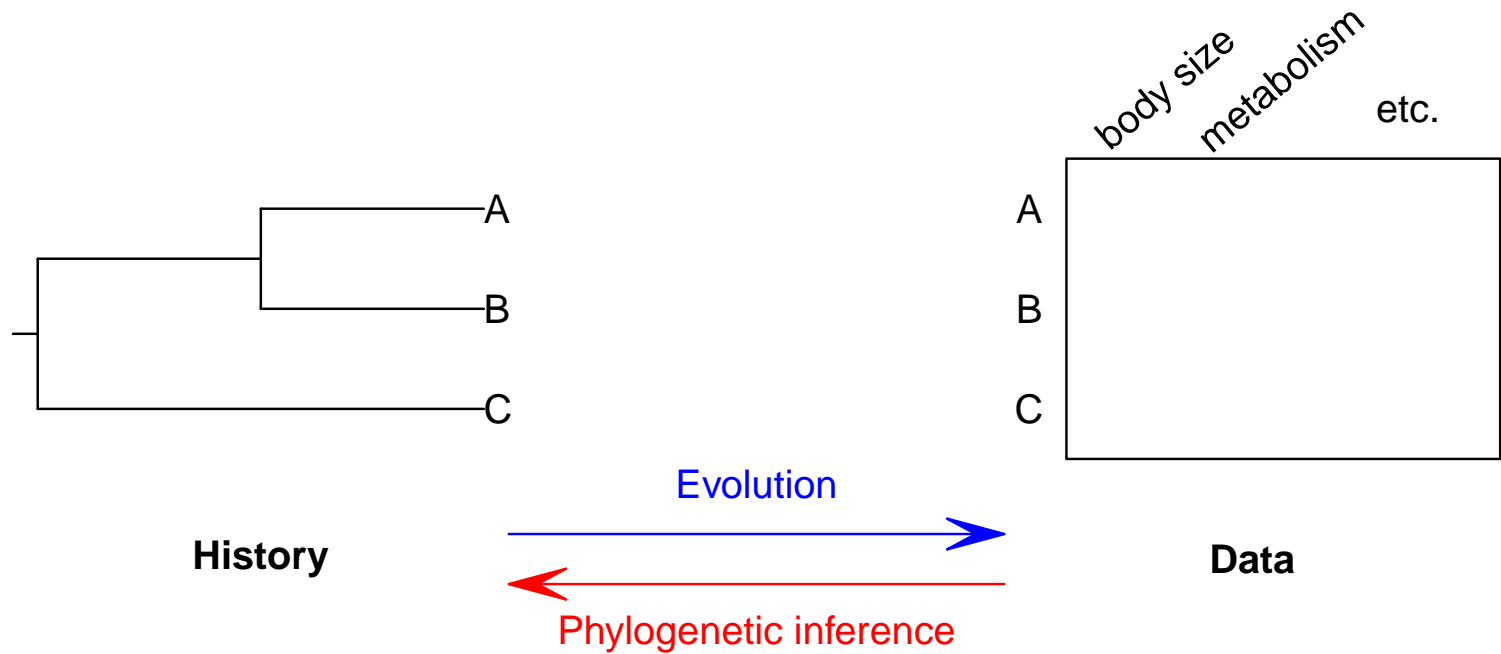
Under the TCL: $SE(\hat{\mu}) = \sqrt{\text{var}(x)/n}$

```
> sqrt(var(x)/length(x))
```

```
[1] 1.118034
```

Why the bootstrap?

- ▶ Small samples: TCL or ML are not accurate.
- ▶ Situations with non-continuous parameters: tree/clustering/phylogeny.

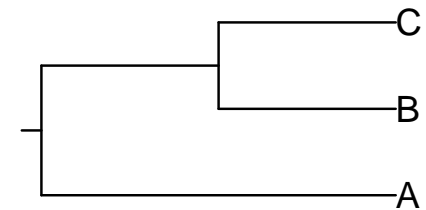
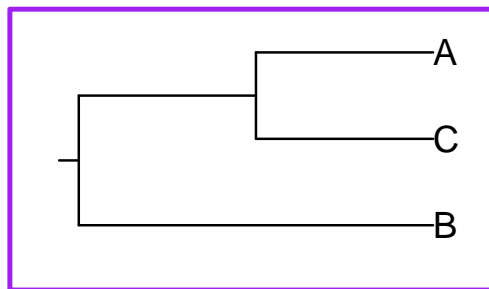
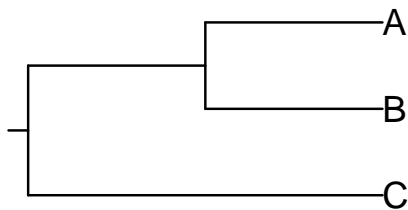


Methods of phylogenetic inference:

- ▶ **Parsimony**: based on the principle of minimum evolution.
- ▶ **Distance**: minimise the discrepancy between the observed distances and those inferred from a tree.

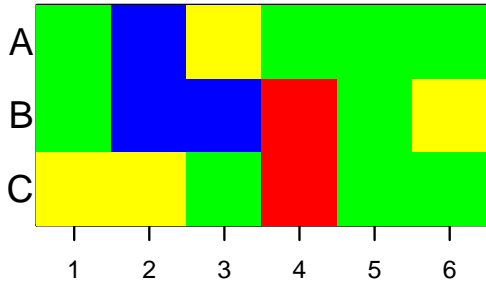
- ▶ **Maximum Likelihood:** Assuming a model of evolution of the characters, maximise the likelihood of the character evolution along the tree.
- ▶ **Bayesian:** id. but maximise the “posterior” probabilities.

These methods help to answer the first fundamental question: ***which tree?***

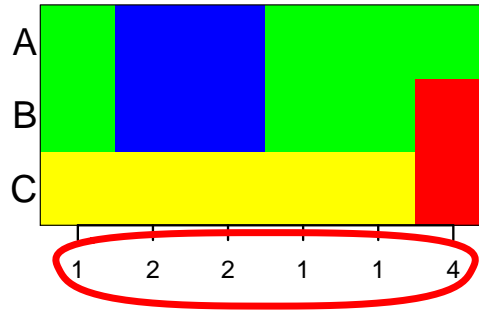


The second fundamental question is: ***how uncertain is this result?***

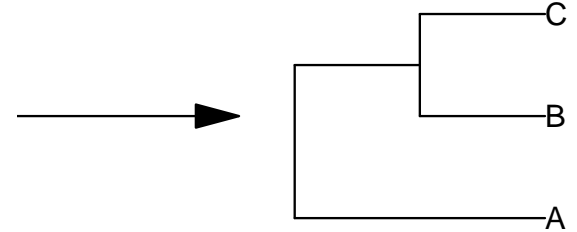
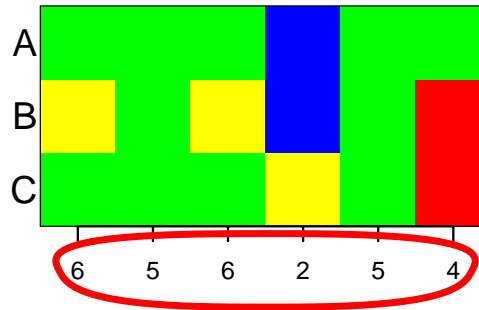
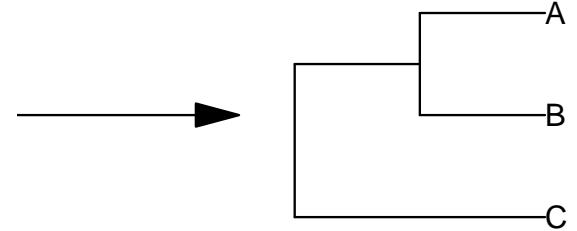
■ A ■ G ■ C ■ T



Bootstrap samples



Bootstrap trees



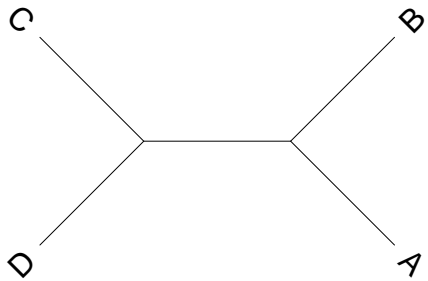
Resampling of columns WITH replacement

⋮

⋮

What statistic to compute with the bootstrap trees?

Each branch of an unrooted tree defines a **split**.



This tree has $n = 4$ tips (or leaves) and thus defines 6 splits:

- ▶ The internal branch (edge) defines one **non-trivial** split: $AB|CD$
- ▶ The terminal branches define four **trivial** splits: $A|BCD$, $B|ACD$, $C|ABD$, $D|ABC$
- ▶ One additional trivial split is defined by the empty set: $\emptyset|ABCD$

An unrooted tree with n tips has $n - 3$ internal branches, and thus defines $2n - 2$ splits: $n - 3$ non-trivial and $n + 1$ trivial.

An unrooted tree with n tips has $n - 3$ internal branches, and thus defines $2n - 2$ splits: $n - 3$ non-trivial and $n + 1$ trivial.

The number of possible splits grows exponentially with n : 2^{n-1}

| | | |
|--------|---|-------------------|
| 4 tips | → | 8 possible splits |
| 10 | | 512 |
| 50 | | 10^{14} |
| 100 | | 10^{29} |

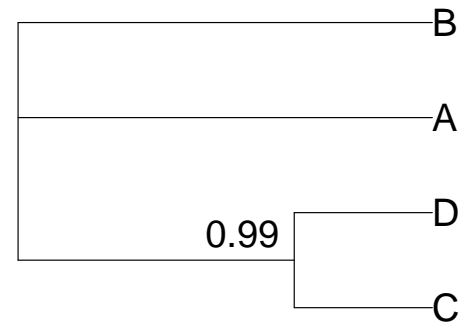
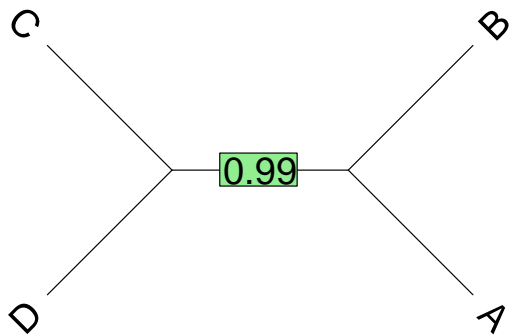
An unrooted tree with n tips has $n - 3$ internal branches, and thus defines $2n - 2$ splits: $n - 3$ non-trivial and $n + 1$ trivial.

The number of possible splits grows exponentially with n : 2^{n-1}

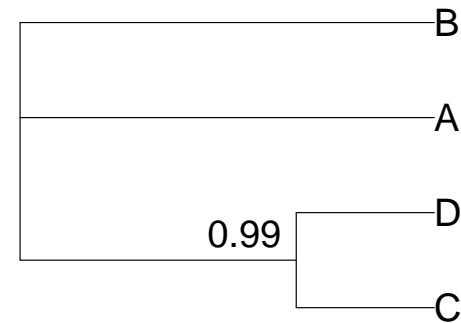
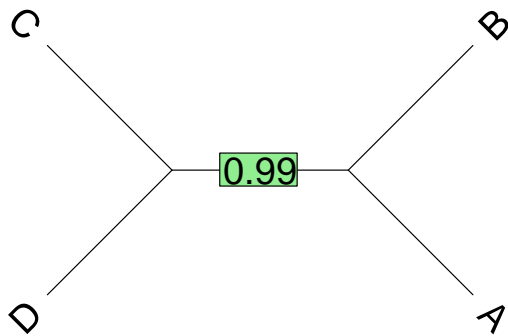
| | | |
|--------|---|-------------------|
| 4 tips | → | 8 possible splits |
| 10 | | 512 |
| 50 | | 10^{14} |
| 100 | | 10^{29} |

The bootstrap statistic in phylogenetics considers the $n - 3$ non-trivial splits and counts how many times they appear in the bootstrap trees. These are the **bootstrap proportions** (BP) and should be interpreted as measures of confidence in the estimated tree (not as probabilities).

The result can be represented graphically with the BP on the internal branch of the estimated tree—usually close to the node defining an MRCA after rooting the tree.



The result can be represented graphically with the BP on the internal branch of the estimated tree—usually close to the node defining an MRCA after rooting the tree.



```
> library(ape)
> data(woodmouse)
> d <- dist.dna(woodmouse)
> tw <- nj(d)
> tw
```

Phylogenetic tree with 15 tips and 13 internal nodes.

Tip labels:

No305, No304, No306, No0906S, No0908S, No0909S, ...

Unrooted; includes branch lengths.

```
> is.rooted(tw)
```

```
[1] FALSE
```

```
> f <- function(x) nj(dist.dna(x))
```

```
> BP <- boot.phylo(tw, woodmouse, f)
```

```
|=====| 100%
```

```
> BP
```

```
[1] 100 51 56 59 67 49 69 74 85 95 86 100 61
```

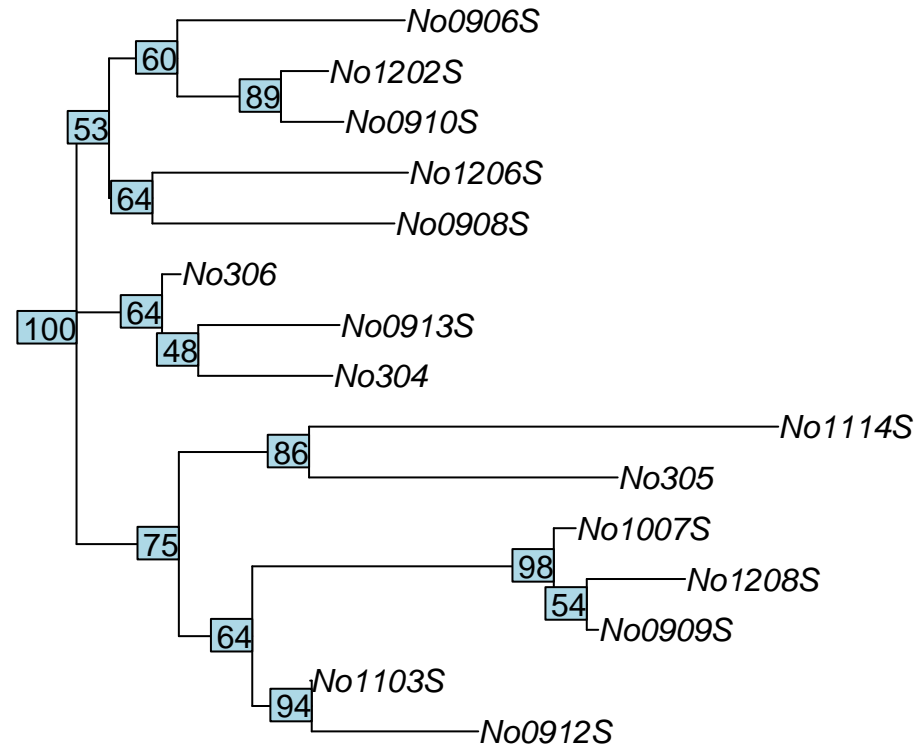
An alternative (though less elegant) is to create the function “on-the-fly”:

```
boot.phylo(tw, woodmouse, function(x) nj(dist.dna(x)))
```

```
boot.phylo(phy, x, FUN, B = 100, block = 1, trees = FALSE,  
          quiet = FALSE, rooted = FALSE)
```

```
> plot(tw)
```

```
> nodelabels(BP, adj = 1)
```





See `?nodeLabels` for all details on how to customise the appearance of the labels (see the examples).

What if the method of phylogenetic inference returns a rooted tree?

```
> fr <- function(x) root(nj(dist.dna(x)), "No1114S")
> twr <- root(tw, "No1114S")
> boot.phylo(twr, woodmouse, fr, rooted = TRUE)
[1] 100  88  65  93 100  56  77  69  57  43  61  65  81
```

In this case `boot.phylo` counts ***clades*** instead of splits.

 It is crucial that the tree is estimated with the same method than used for the bootstrap.

So far we have used a neighbour-joining (NJ) method with distances calculated with Kimura's two-parameter model (K80). What about other methods? For instance, BIONJ method with Tamura's (1993) distance:

```
f <- function(x) bionj(dist.dna(x, "T93"))
```

Euclidean distance and UPGMA (requires the package phangorn):

```
f <- function(x) upgma(dist(x))
```

Maximum likelihood with molecular sequences:

```
> library(phangorn)
> o <- optim.pml(pml(tw, as.phyDat(woodmouse)))
> ctr <- pml.control(trace = 0)
```

```
> TR <- bootstrap.pml(o, control = ctr)
> TR
100 phylogenetic trees
```

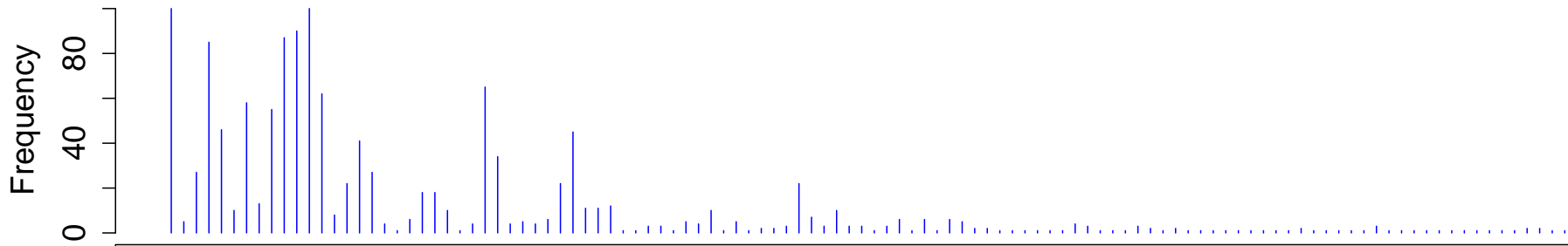
The bootstrap proportions can then be calculated with `prop.clades`.

The bootstrap proportions refer only to the splits observed in the estimated tree.
What about the other splits?

```
> res <- boot.phylo(tw, woodmouse, f, trees = TRUE, quiet = TRUE)
> res
$BP
 [1] 100  49  50  52  63  46  73  61  86  93  90 100  62

$trees
100 phylogenetic trees

> pp <- prop.part(res$trees)
> plot(pp)
```

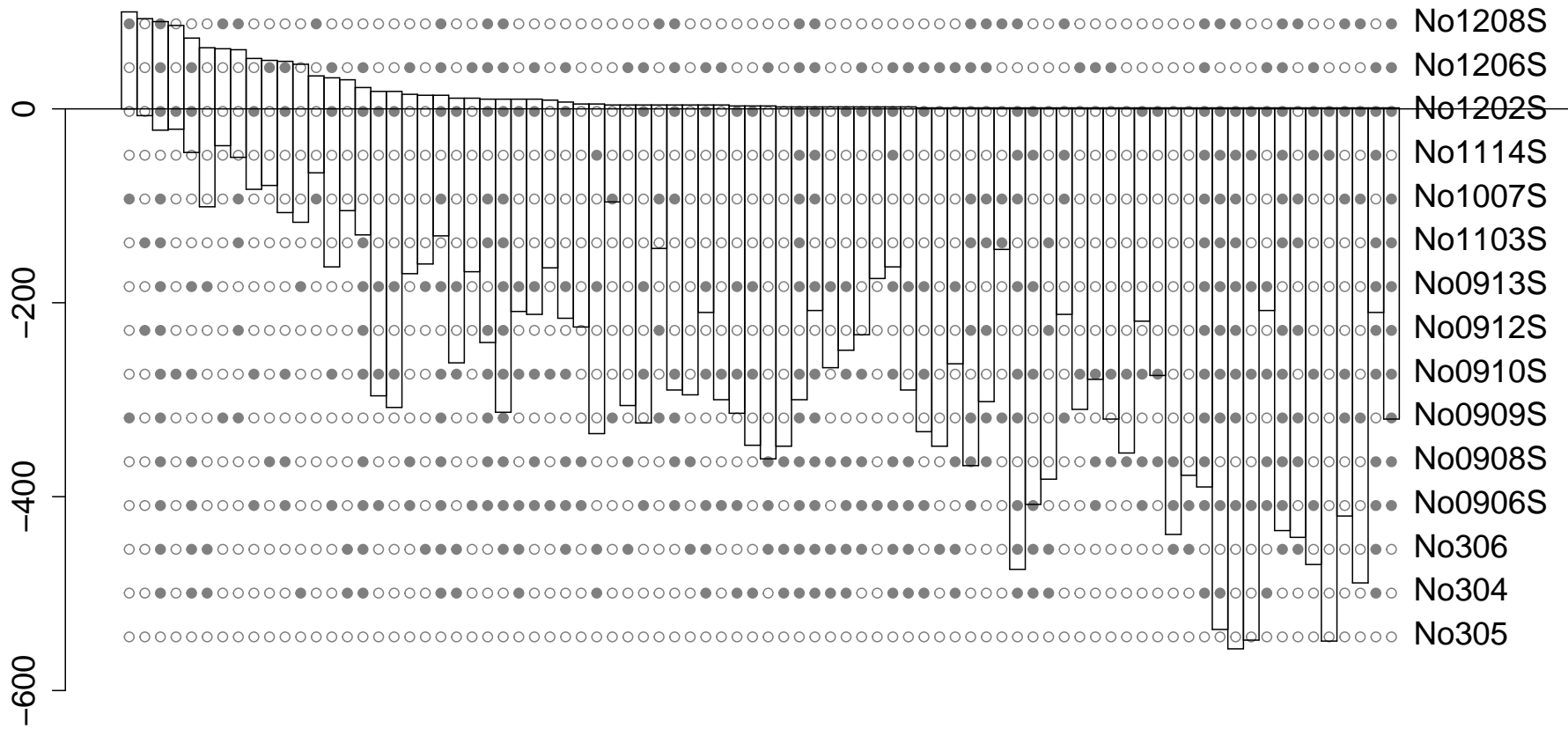


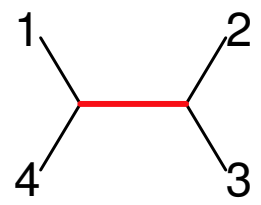
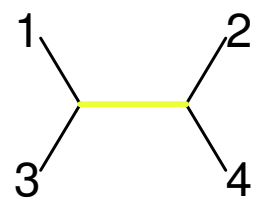
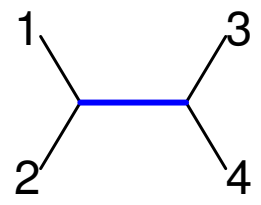
| | | | | | | | | | | | | | | | | | | | | | | | | |
|---------|-----|------|-----|-----|-----|----|-----|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| No1208S | o o | o oo | o | ooo | ooo | o | o | oooo | oo | oo | oooo | o | o | oooo | oooo | oo | oooo | ooo | oo | o | oo | oo | o | |
| No1206S | oo | o | oo | o | o | o | o | oo | oo | oooo | o | oo | o | oo | oo | o | oo | oo | oooo | o | oo | oo | oo | |
| No1202S | o | o | ooo | oo | o | oo | o | oo | o | oo | oooo | oooo | o | oo | oooo | oooo | oo | oooo | ooo | oooo | oooo | oooo | oooo | |
| No1114S | o | o | o | o | o | o | oo | o | o | oooo | o | oo | oooo | oooo | ooo | oooo | oo | oooo | ooo | oooo | oooo | o | oo | oooo |
| No1007S | o | o | o | o | o | oo | ooo | o | o | oooo | oo | oo | oooo | o | o | oooo | oooo | oo | oooo | ooo | oo | oo | oo | o |
| No1103S | o | o | oo | o | o | o | o | oooo | oo | oooo | o | o | oooo | oo | o | oo | oooo | ooo | oo | oo | oo | oo | oo | o |
| No0913S | o | o | ooo | o | oo | o | oo | o | ooo | o | oo | oo | ooo | oooo | oo | oooo | oooo | o | oo | oooo | oooo | oooo | oooo | oo |
| No0912S | o | o | oo | o | o | o | o | oooo | oo | oooo | o | o | oo | oo | o | oo | oooo | ooo | oo | oo | oo | oo | oo | o |
| No0910S | o | o | ooo | oo | o | oo | o | oo | o | oo | oooo | ooo | oo | oooo | oooo | oo | oooo | ooo | oo | oooo | oooo | oooo | oooo | oo |
| No0909S | o | o | o | oo | o | o | o | ooo | o | o | oooo | oo | oo | oooo | o | o | oooo | oooo | oo | oooo | ooo | oo | oo | o |
| No0908S | oo | o | oo | o | o | o | o | o | oooo | oo | oo | o | ooo | o | o | o | oooo | o | oo | oooo | oooo | oooo | oooo | oo |
| No0906S | oo | ooo | o | o | oo | o | oo | oo | oooo | oo | oooo | oo | oooo | oo | oooo | oo | oooo | ooo | oo | oooo | oooo | oooo | oooo | oo |
| No306 | o | o | ooo | o | o | o | o | oo | o | oo | oooo | ooo | oo | o | oo | ooo | oo | o | oo | o | o | o | o | o |
| No304 | o | o | oo | o | o | oo | oo | o | ooo | o | oo | oo | ooo | oooo | oo | ooo | oo | oo | oooo | oo | o | o | o | oo |
| No305 | o | o | o | o | o | o | o | oo | o | o | oooo | oo | oooo | oo | o | ooo | o | oo | oooo | oooo | oo | oooo | oooo | oo |

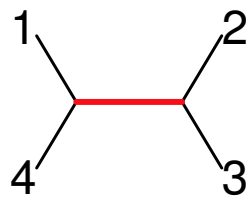
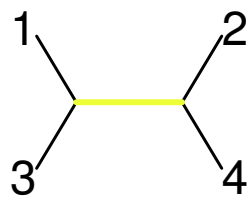
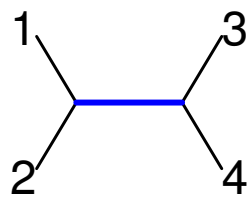
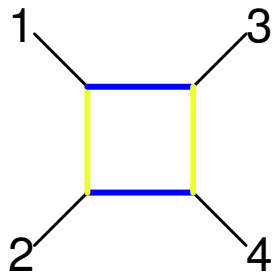
- ▶ **support**: the observed frequency of the split;
- ▶ **conflict**: the sum of the support values of the splits that are not compatible with the considered one.

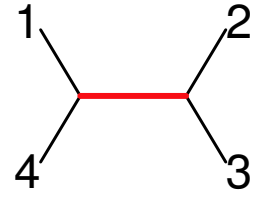
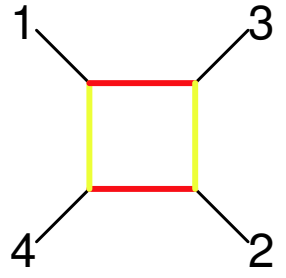
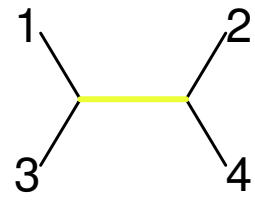
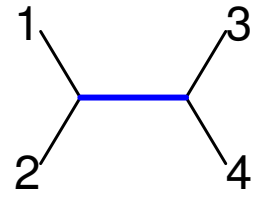
Two splits are not compatible if they cannot be observed in the same tree (e.g., $AB|CD$ and $A|BCD$ are compatible, while $AB|CD$ and $AC|BD$ are not).

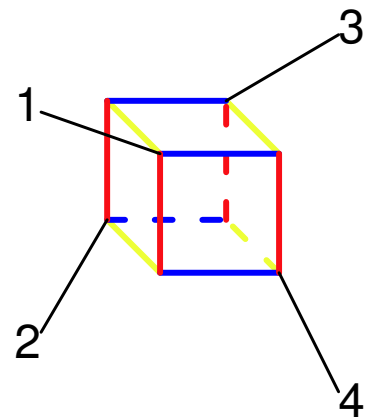
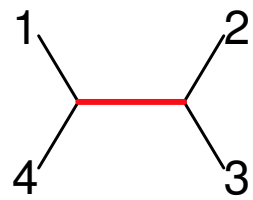
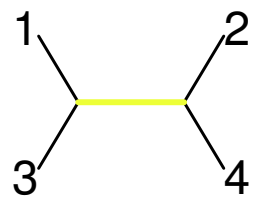
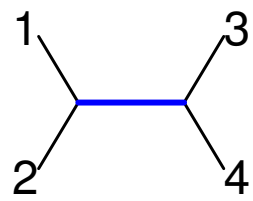
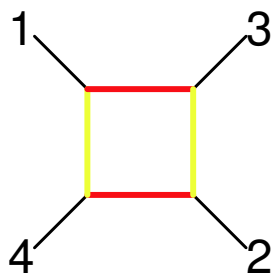
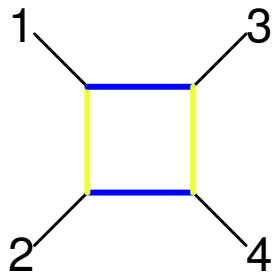
Lento plot





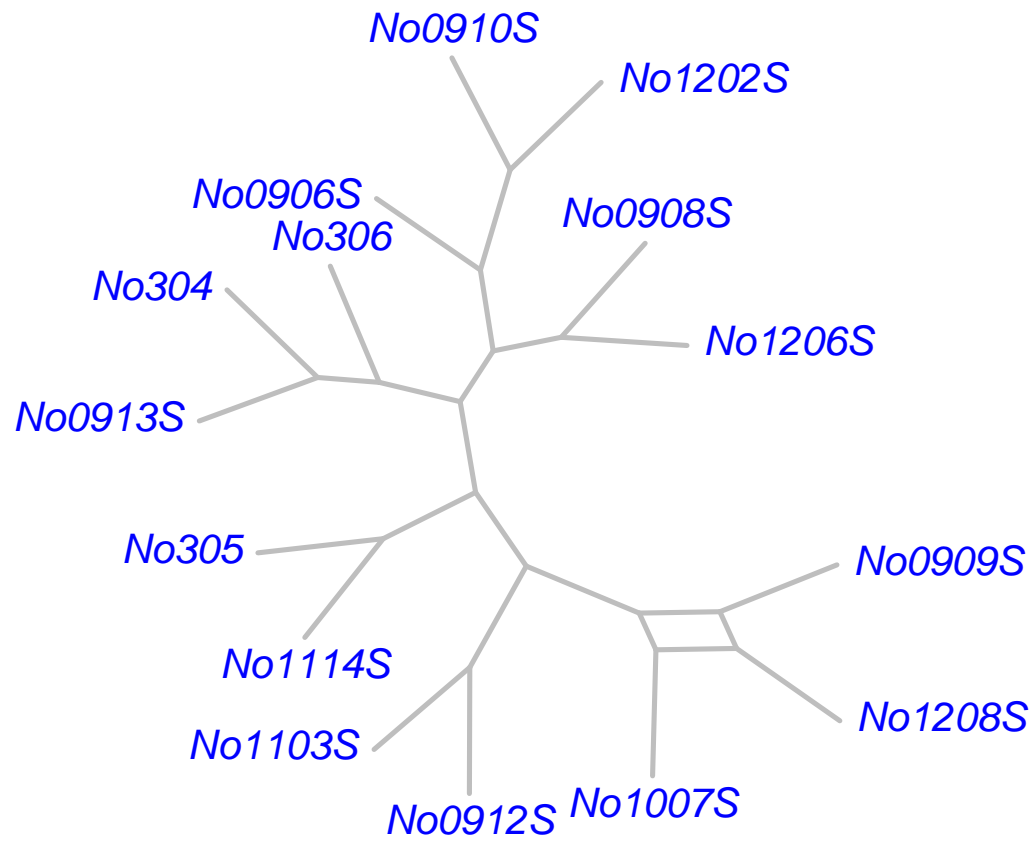






Consensus network:

```
> CN <- consensusNet(res$trees)
> plot(CN, "2")
```



The bootstrap is used to assess uncertainty (or confidence) of a ***single*** phylogeny.

How to compare different phylogenies?

The bootstrap is used to assess uncertainty (or confidence) of a ***single*** phylogeny.

How to compare different phylogenies?

A special test based on the bootstrap, the Shimodaira–Hasegawa test, must be used. The principle is to perform a statistical comparison based on resampling the likelihood of the trees.

The bootstrap is used to assess uncertainty (or confidence) of a **single** phylogeny.

How to compare different phylogenies?

A special test based on the bootstrap, the Shimodaira–Hasegawa test, must be used. The principle is to perform a statistical comparison based on resampling the likelihood of the trees.

H_0 : the differences in likelihood are not statistically different

H_1 : the tree with the largest likelihood is statistically better

```
tr2 <- bionj(d)
```

```
X <- as.phyDat(woodmouse)
```

```
fit0 <- optim.pml(pml(tr2, X))
```

```
fit2 <- optim.pml(pml(tr2, X))  
SH.test(fit0, fit2)
```

See `?SH.test` for another example.

Summary:

- ▶ The bootstrap is a very general method to assess confidence in statistical estimation.
- ▶ Bootstrap proportions assess confidence in phylogeny estimation by counting splits.
- ▶ Complementary methods (Lento plot, consensus networks) help to explore alternative splits not observed in the estimated tree.
- ▶ The Shimodaira–Hasegawa test must be used to test statistically different trees with the bootstrap.