8.xi.

1. evaluating results

2. mono-, para- & polyphyly

3. various optimality criteria

4. summary

Evaluating results
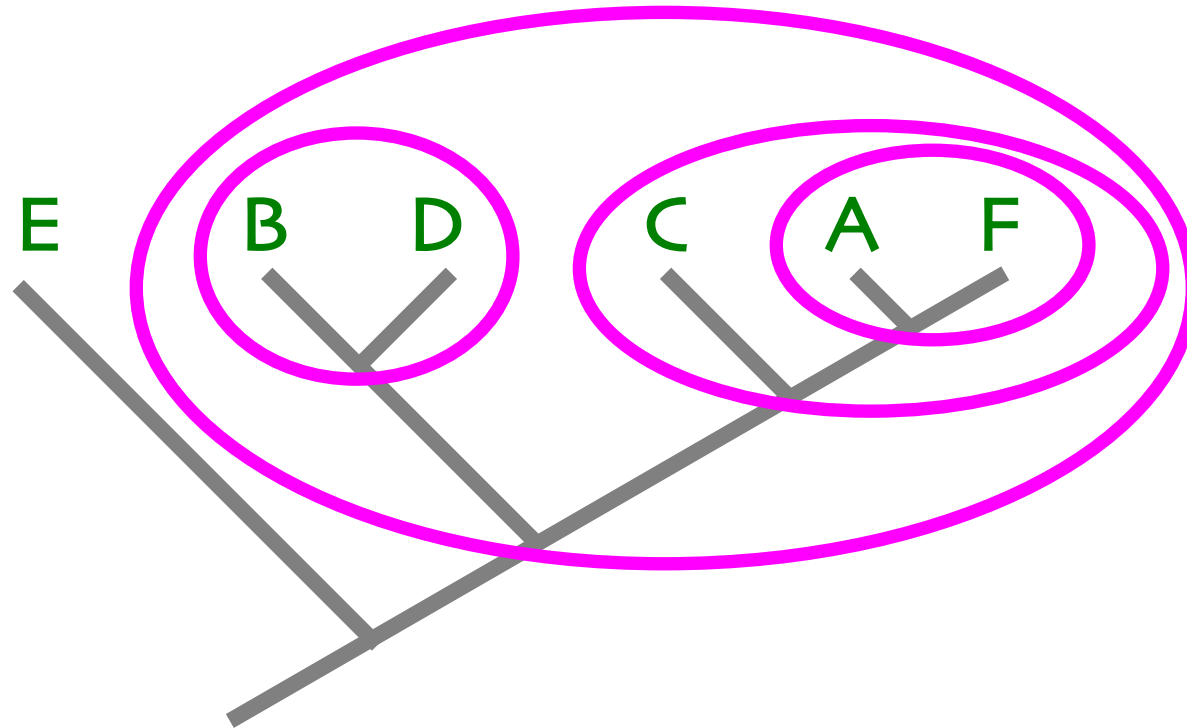
can we trust the results obtained?
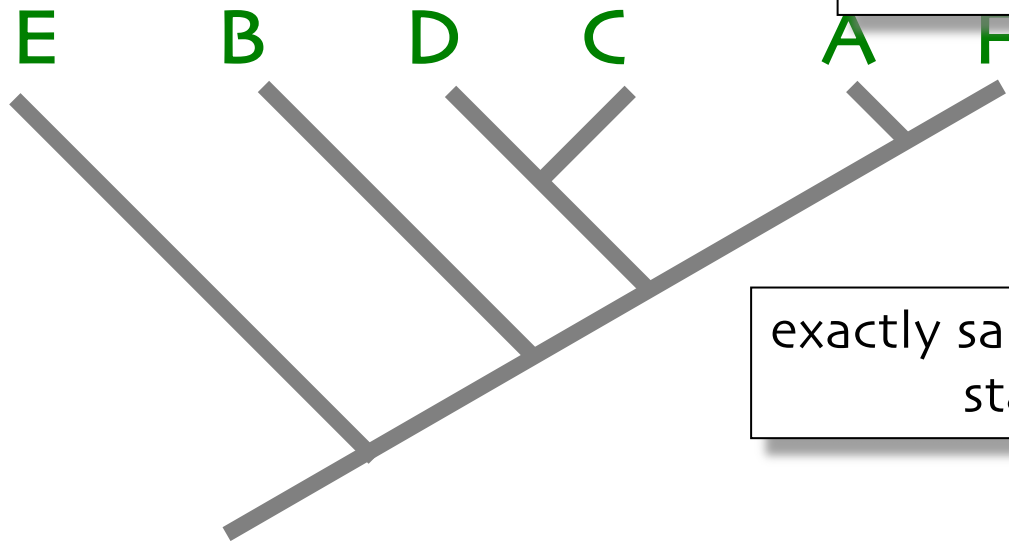
are part of the results simply accidental?

which PARTS of tree are most reliable?

Clades WITHIN clades
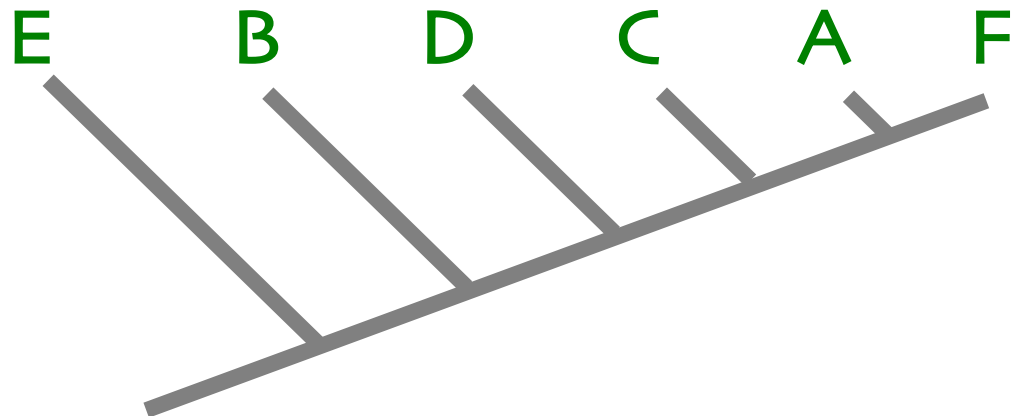
Numerous equally parsimonious trees

EQUALLY simple explanations
for data

E B D C A F

exactly same number of ch.
state changes

E B D C A F
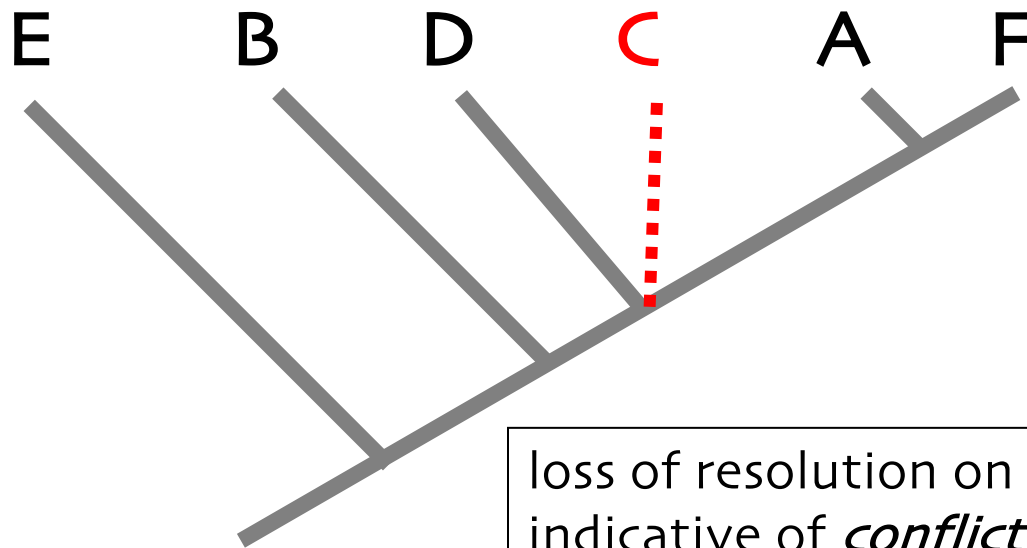
Consensus tree

loss of resolution on consensus is indicative of *conflict* in characters

Evaluating results

support values

3 commonly used methods:

Bremer support value

branch support

Bootstrap

Parsimony jackknifing

Evaluating results
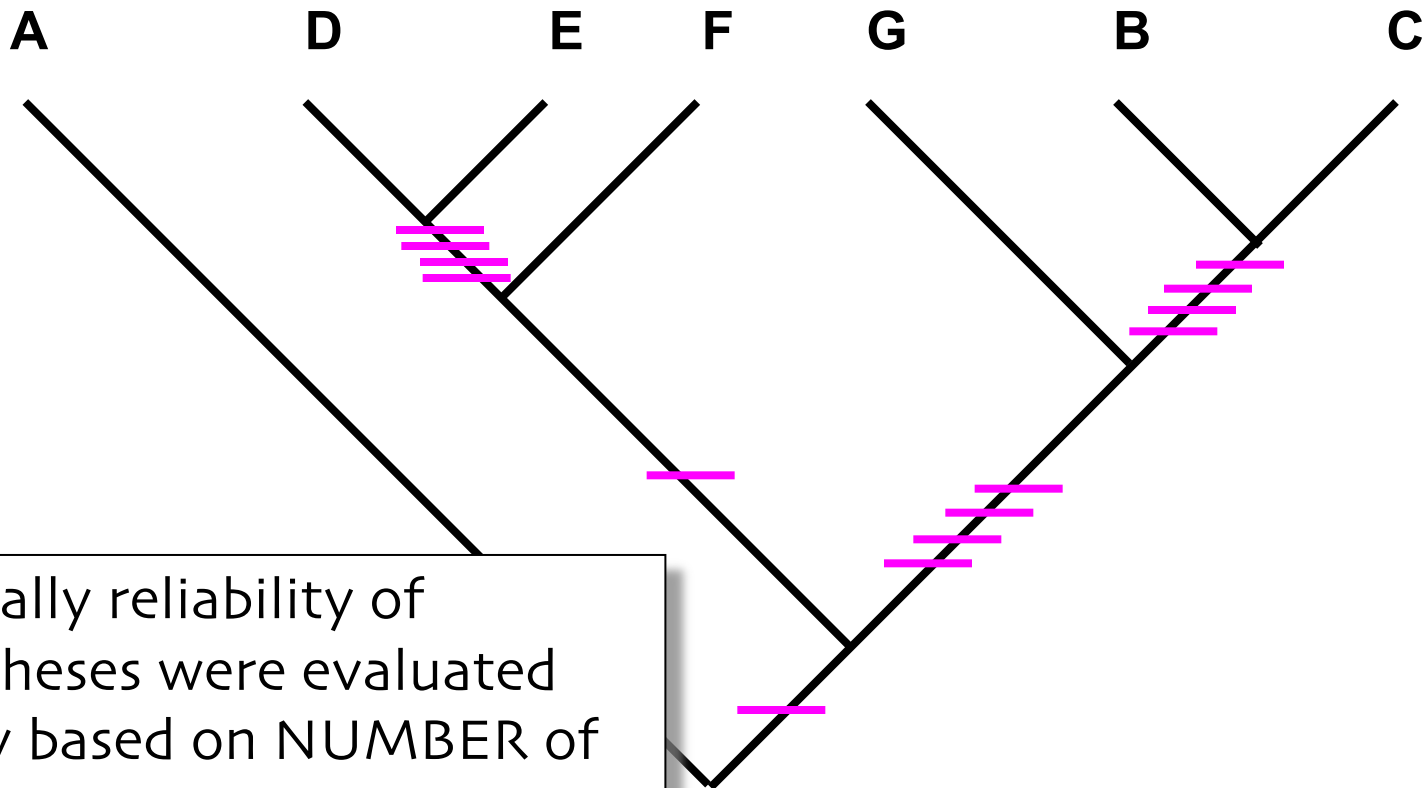
3 commonly used methods:

Bremer support value

Bootstrap

Parsimony jackknifing
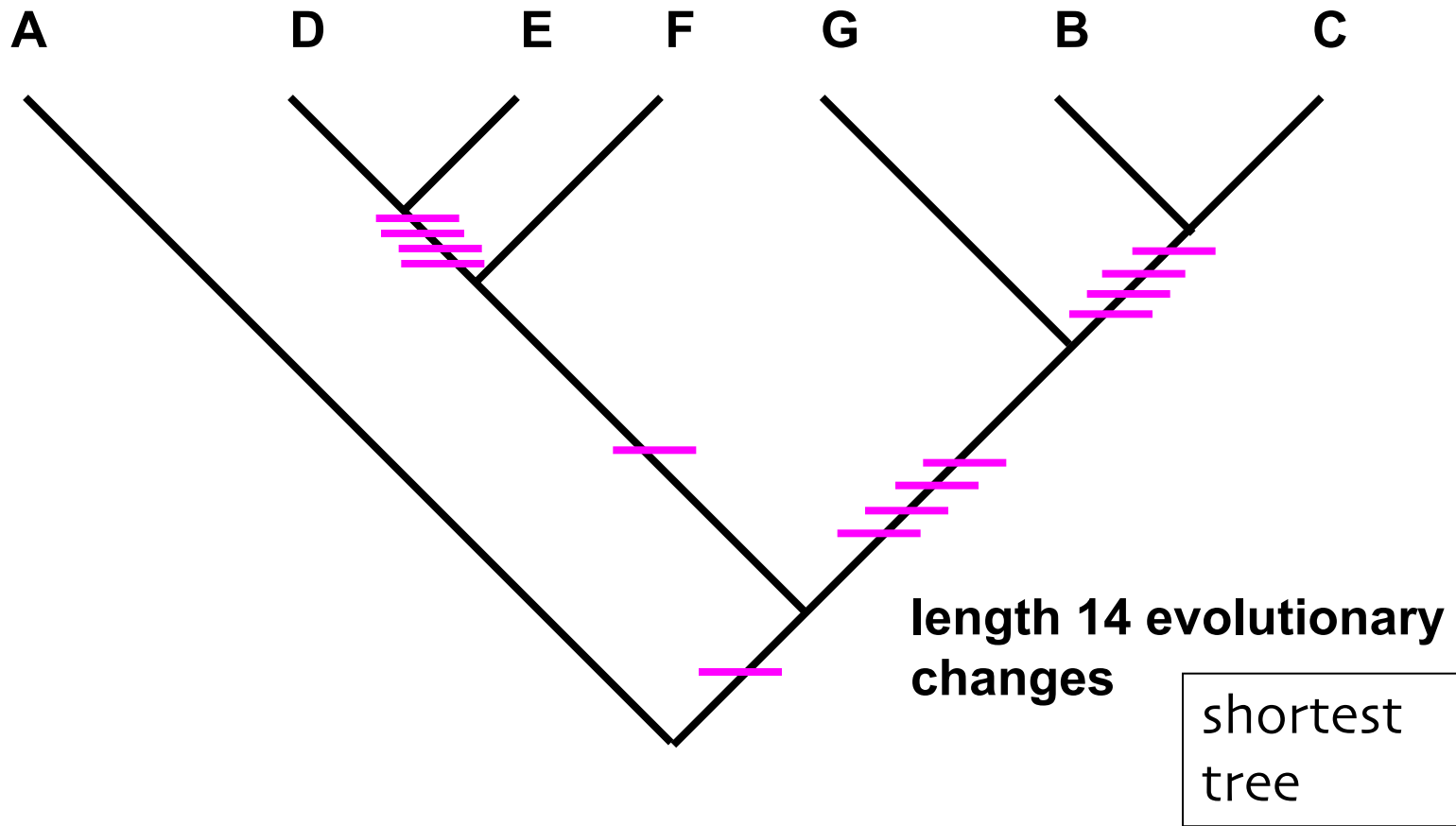
Bremer support

originally reliability of hypotheses were evaluated purely based on NUMBER of characters supporting branch

Bremer support

```
       0 1 2 3 4 5 6 7 8 9

A      0 0 0 0 0 0 0 0 0 0          7 X 10
B      1 0 1 1 1 1 1 1 1 1
C      1 0 1 1 1 1 1 1 1 1
D      1 1 0 0 0 0 1 1 1 1
E      1 1 0 0 0 0 1 1 1 1
F      1 1 0 0 0 0 0 0 0 0
G      1 0 1 1 1 1 0 0 0 0
```
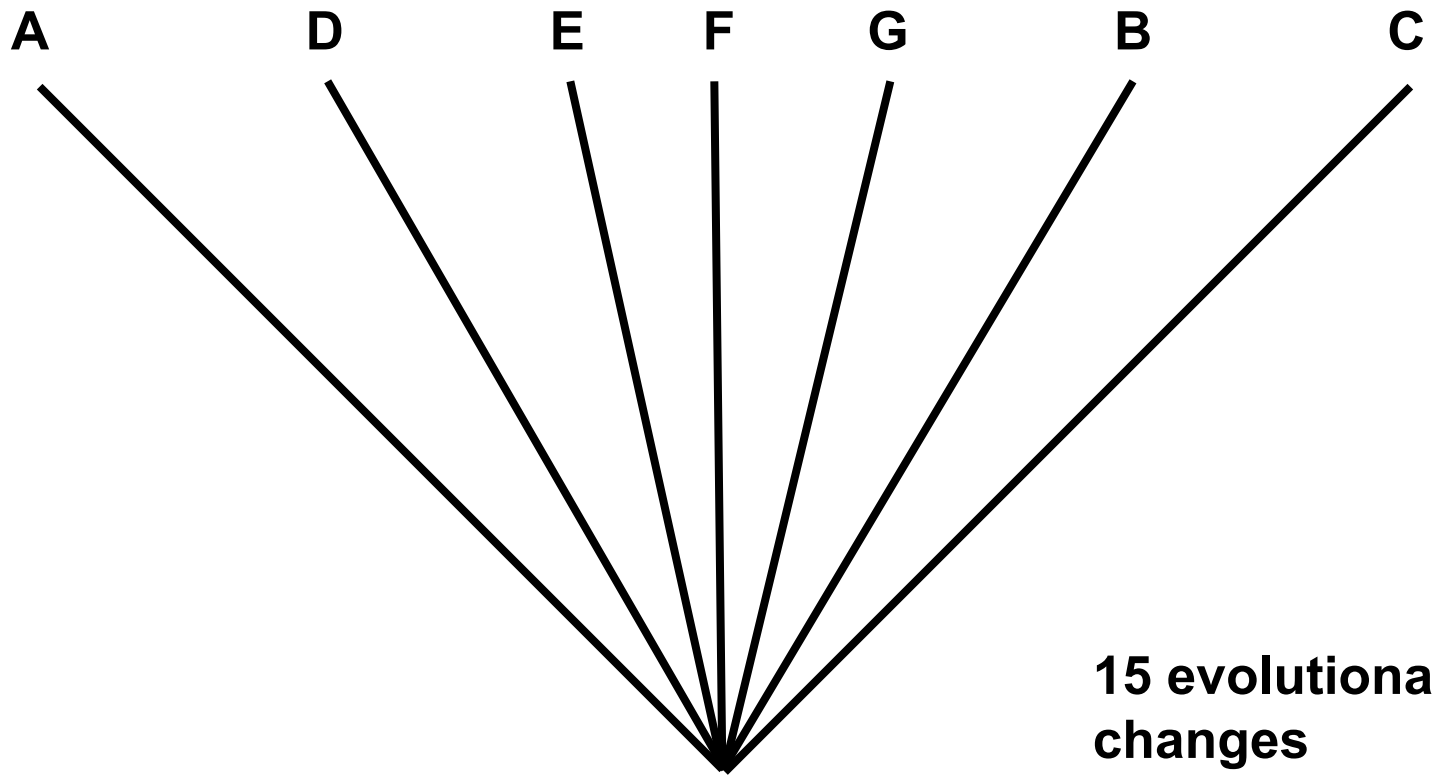
Bremer support

length 14 evolutionary changes

shortest tree

Bremer support

A   D   E   F   G   B   C

15 evolutionary changes

consensus of trees with length ≤ 15

TAXONOMIC CHARACTERS

characters used in phylogenetic analyses are assumed to be INDEPENDENT of other characters
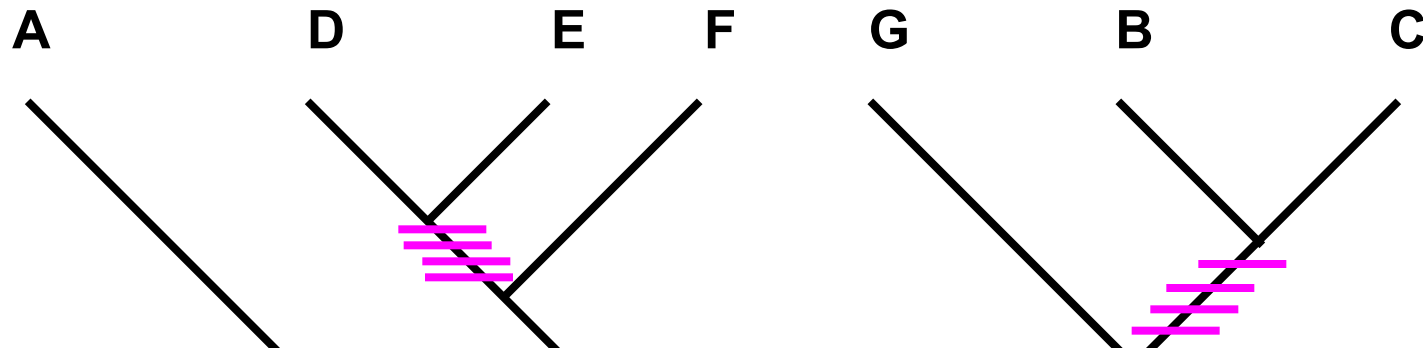
NO genetic correlation

ALL these considered to be equally valuable =
   potentially useful for phylogenetic analyses

but characters DO INTERACT within the matrix

character congruence

Bremer support

A   D   E   F   G   B   C

Goodman, M. & al. 1982. New perspectives in the molecular biological analysis of mammalian phylogeny. *Acta Zool. Fennica* 169: 19-35.

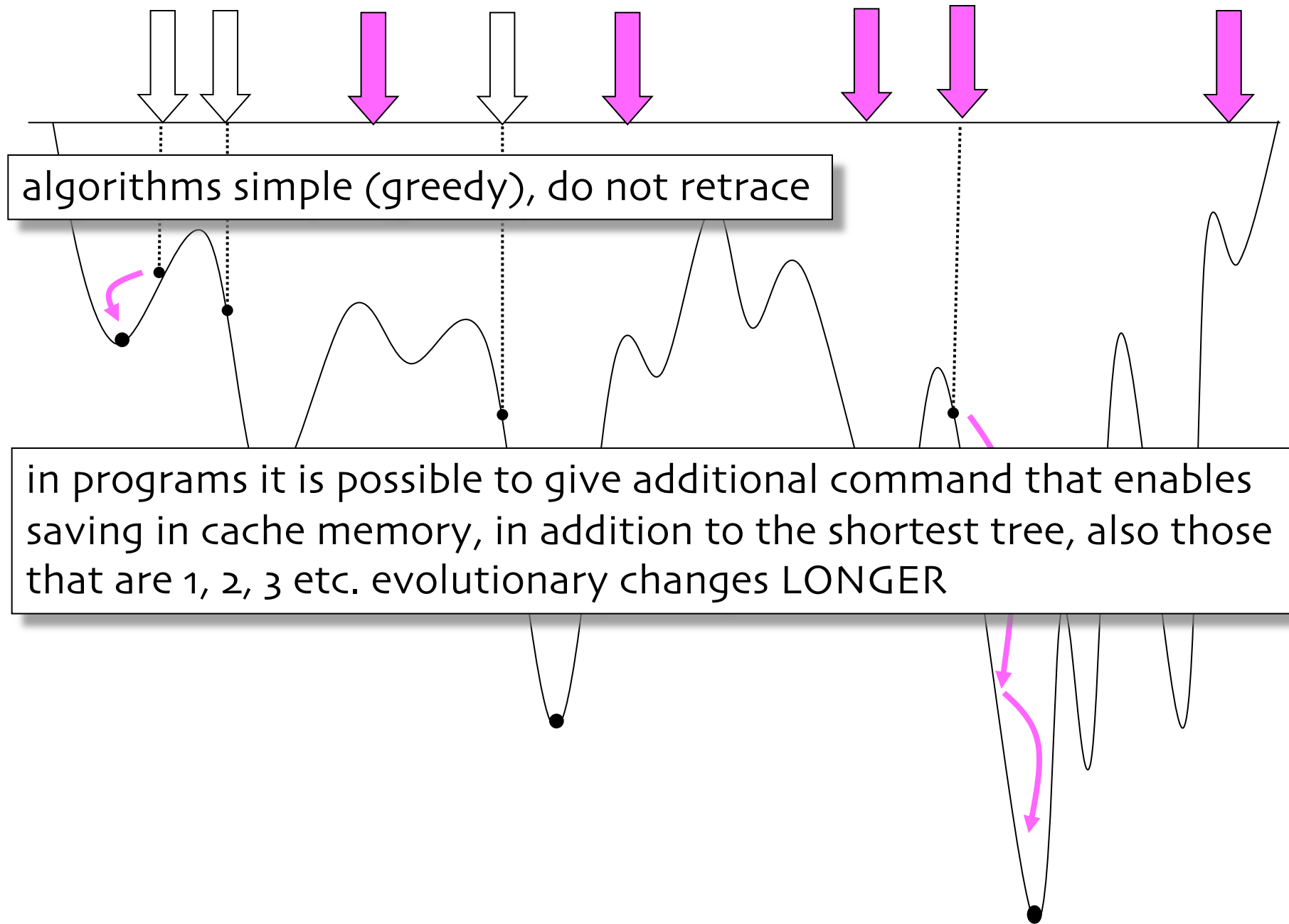Goodman-Bremer support

decay index

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

Bremer support

1. search for shortest tree

2. consensus groups disappearing at this stage
with Bremer support value = 0

3. new search for trees with length L $\leq$ L+1
(L = length of shortest tree found so far),
consensus of these for these Bremer support value = 1

algorithms simple (greedy), do not retrace

in programs it is possible to give additional command that enables
saving in cache memory, in addition to the shortest tree, also those
that are 1, 2, 3 etc. evolutionary changes LONGER

L+3

L+2

L+1

for higher Bremer support values it is more likely that parts of the tree space **REMAIN UNEXPLORED**

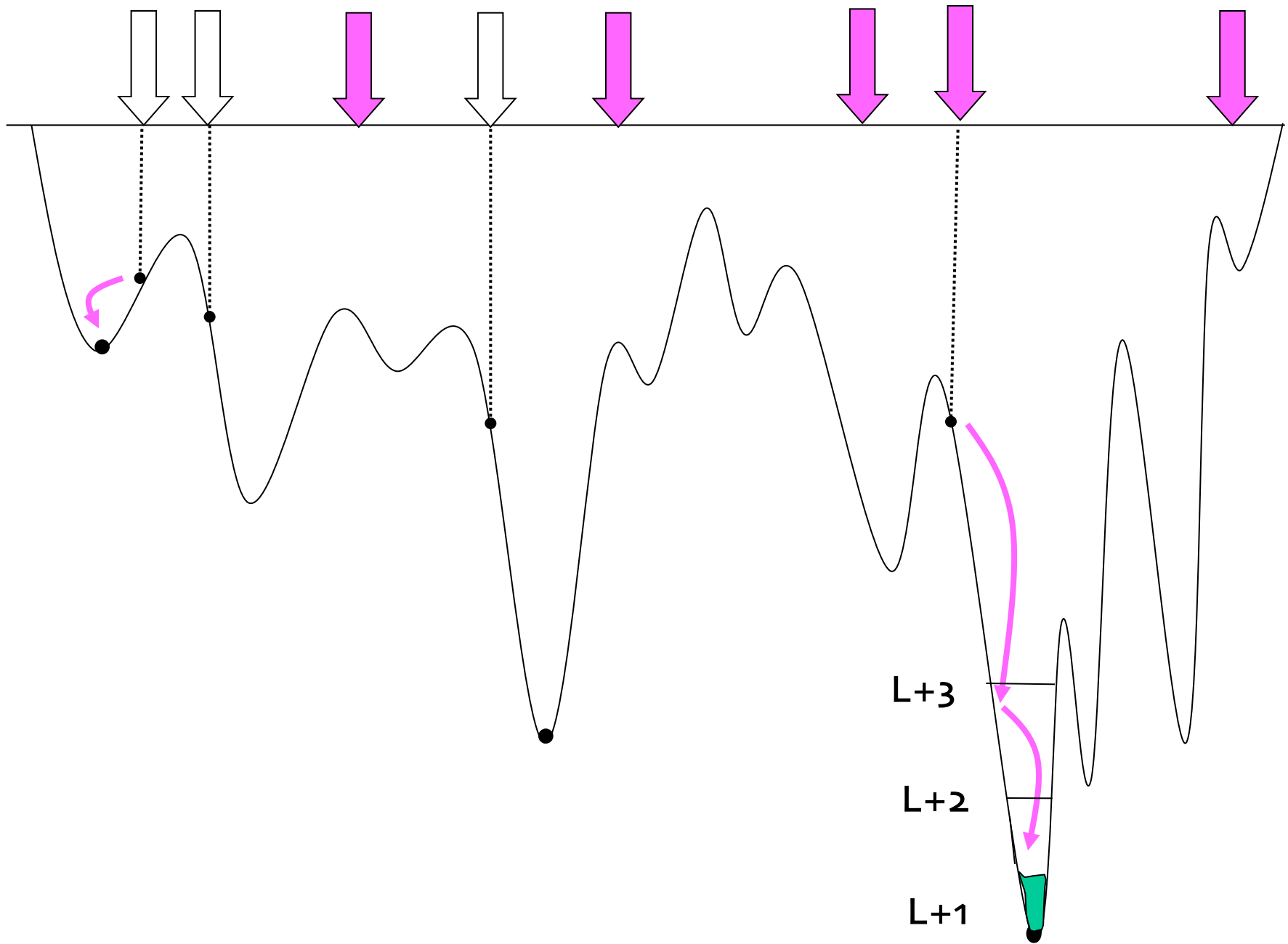Bremer support values LARGER than actual values
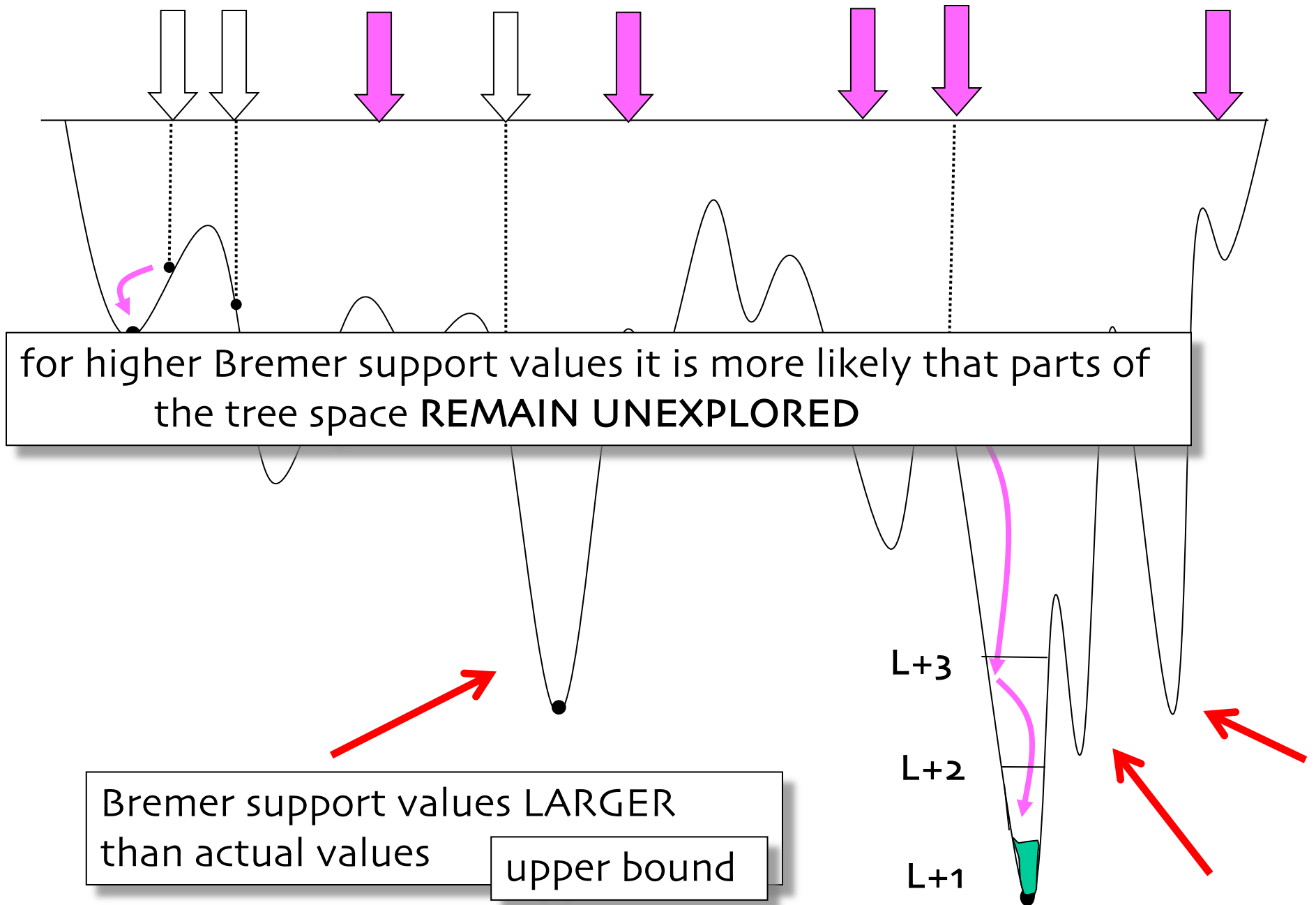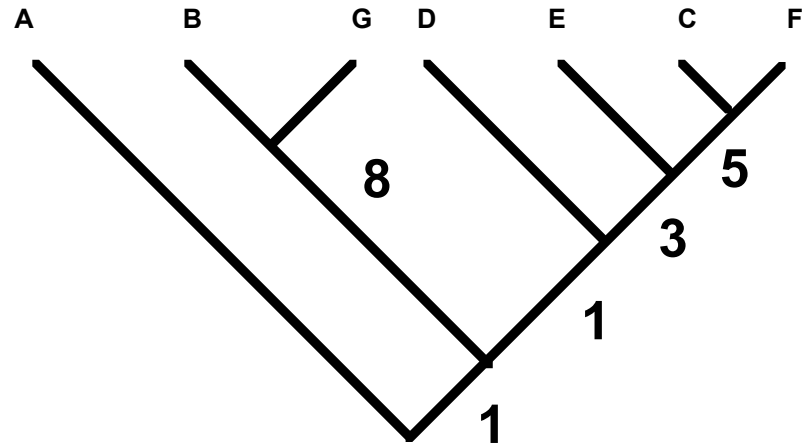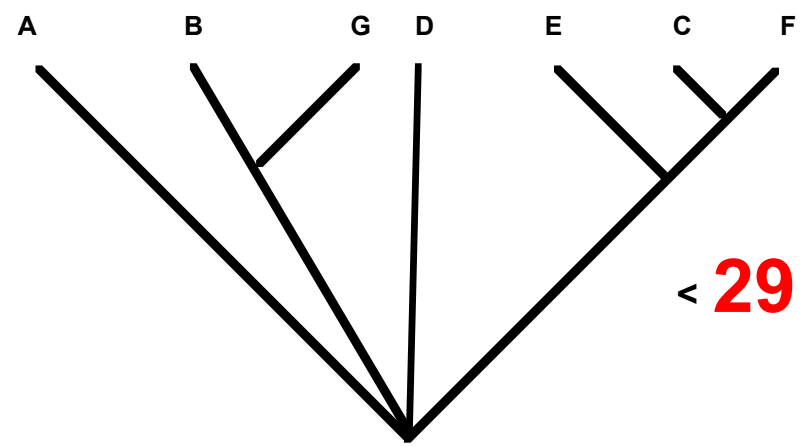
upper bound

L+3

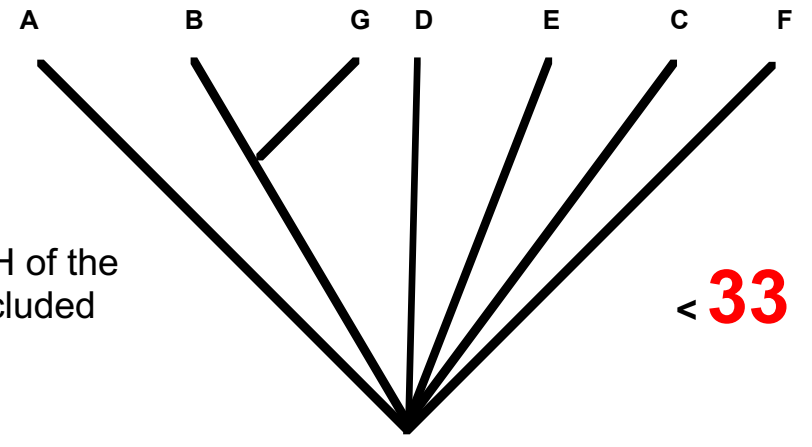L+2

L+1

Bremer support

1. search for shortest tree

2. consensus | groups disappearing at this stage with Bremer support value = 0

3. new search for trees with length L $\leq$ L+1
    (L = length of shortest tree found so far),
    consensus of these | for these Bremer support value = 1

4. continued until consensus has lost ALL
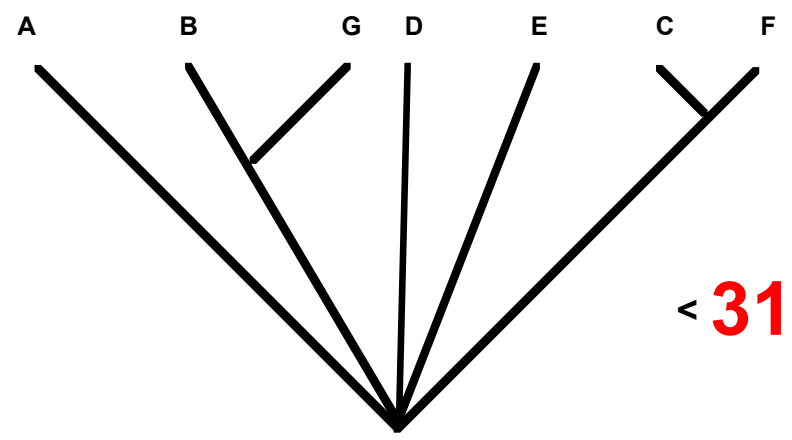   resolution (only polytomy remains)

A B G D E C F

8

5

3

1

1

28 LENGTH of the shortest tree

A B G D E C F

< **29** LENGTH of the trees included

A B G D E C F

< **33**

A B G D E C F

< **31**

A B G D E C F

< **36**

Bremer support

A   B G   D   E   C   F
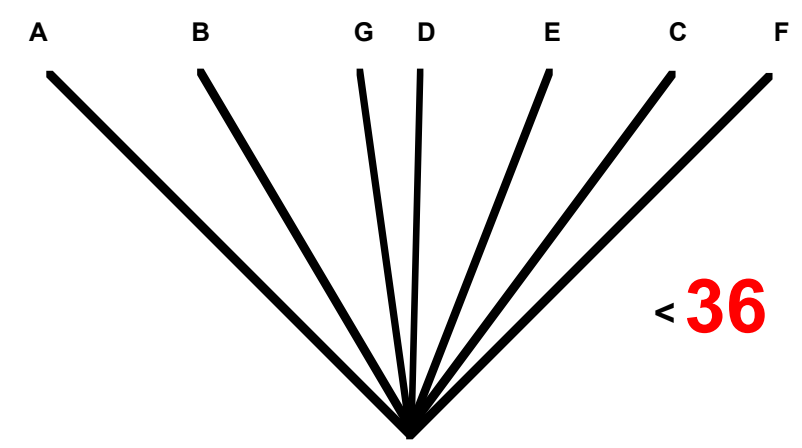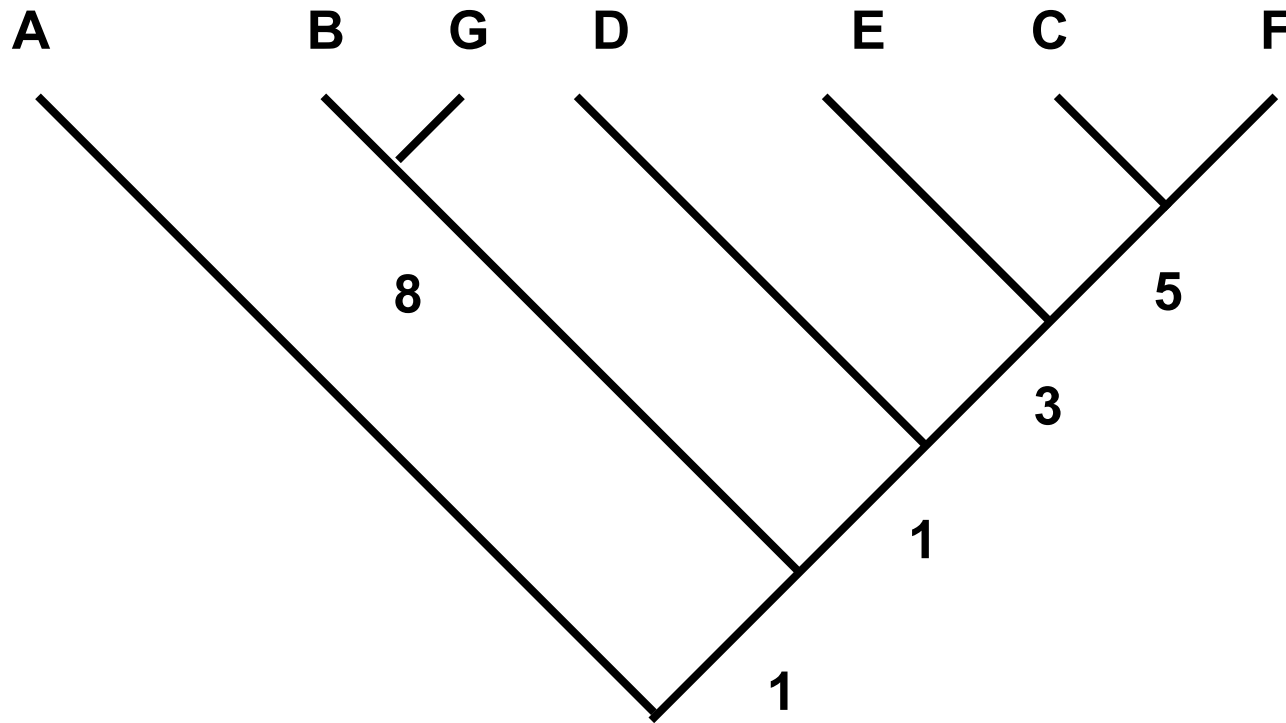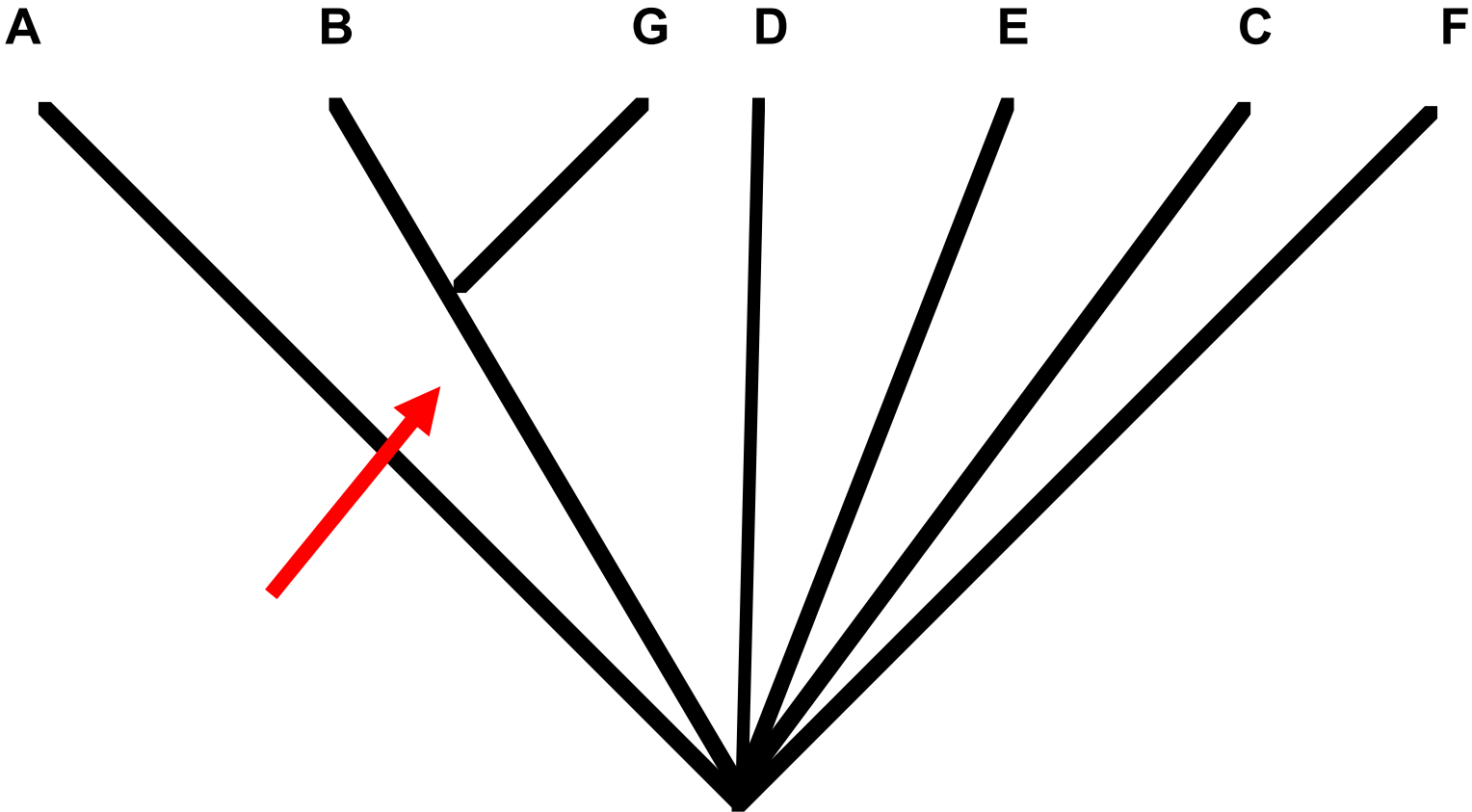
8

5

3

1

1

18
Bremer support value for whole tree

Bremer support

Bremer support value can be calculated also for *individual parts of tree* using constrained search

this is done by searching for shortest tree using as a *constraint* such a tree that includes ONLY the group for which we want to know support value

constraint tree is otherwise totally without resolution (unresolved)

Bremer support

Bremer support value can be calculated also for *individual parts of tree* using constrained search

this is done by searching for shortest tree using as a <u>*constraint*</u> such a tree that includes ONLY the group for which we want to know support value

constraint tree is otherwise totally without resolution (unresolved)

search for trees that are in CONFLICT with the constraint tree
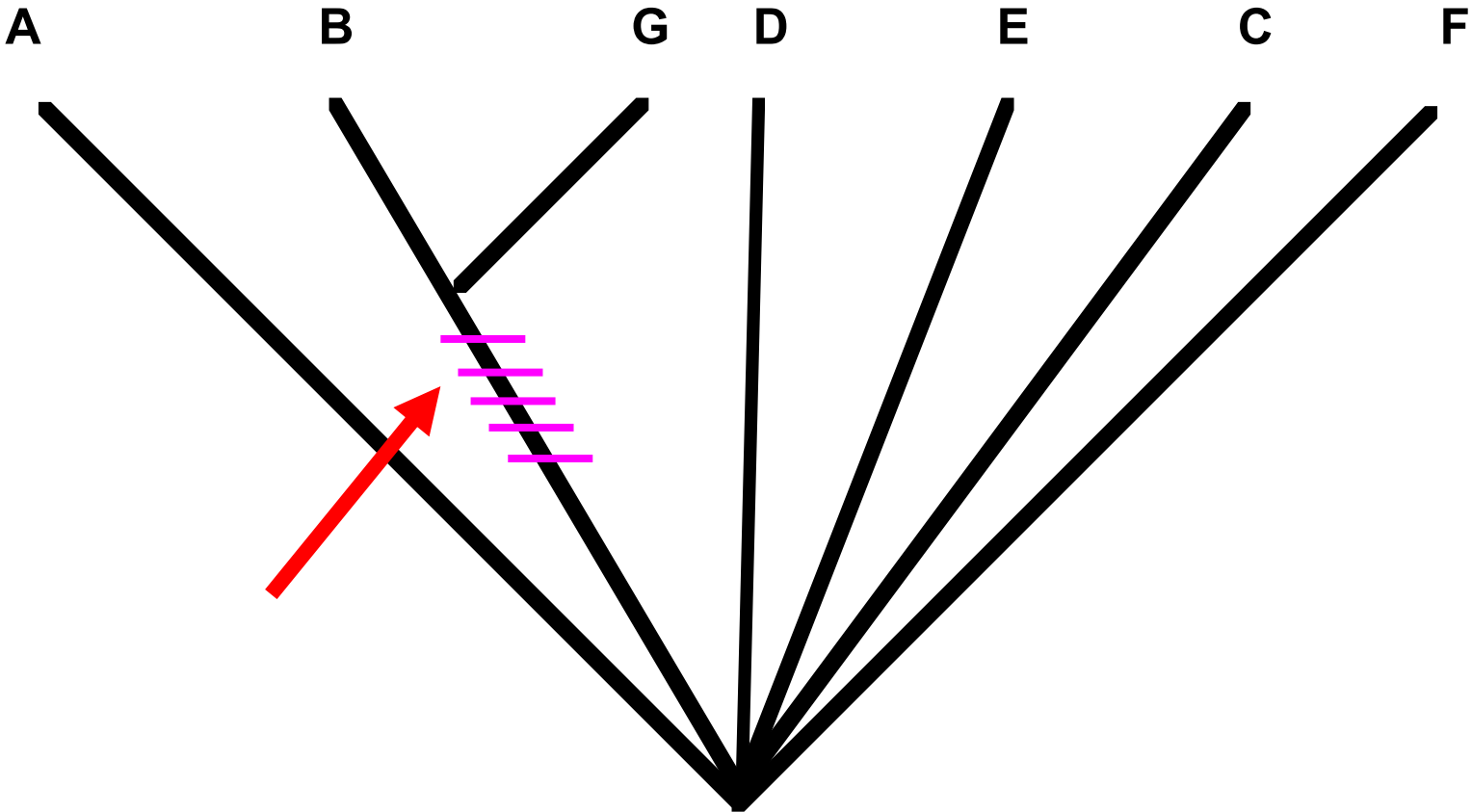
Bremer support

Bremer support value can be calculated also for *individual parts of tree* using constrained search

this is done by searching for shortest tree using as a _constraint_ such a tree that includes ONLY the group for which we want to know support value

constraint tree is otherwise totally without resolution (unresolved)

search for trees that are in CONFLICT with the constraint tree

support value for the group is the difference between length of the shortest tree obtained without constraint and the one found by using constraint

Evaluating results

3 commonly used methods:

Bremer support   Relative Fit Difference (RFD)

Goloboff, P. & Farris, J.S. 2001. Methods for quick
    consensus estimation. *Cladistics* 17: S26-S34.

Parsimony jackknifing

Relative Fit Difference

$$RFD = \frac{F - C}{F}$$

F = synapomorphies of the group inspected

C = synapomorphies of groups in CONFLICT with the group inspected

$$0 < RFD < 1$$

ability to distinguish between characters that have the same Bremer support value

e.g. F = 5, C = 0 vs. F= 100, C= 95 same Bremer support value

RFD values 1 & 0,053

Evaluating results

3 commonly used methods:

Bremer support value
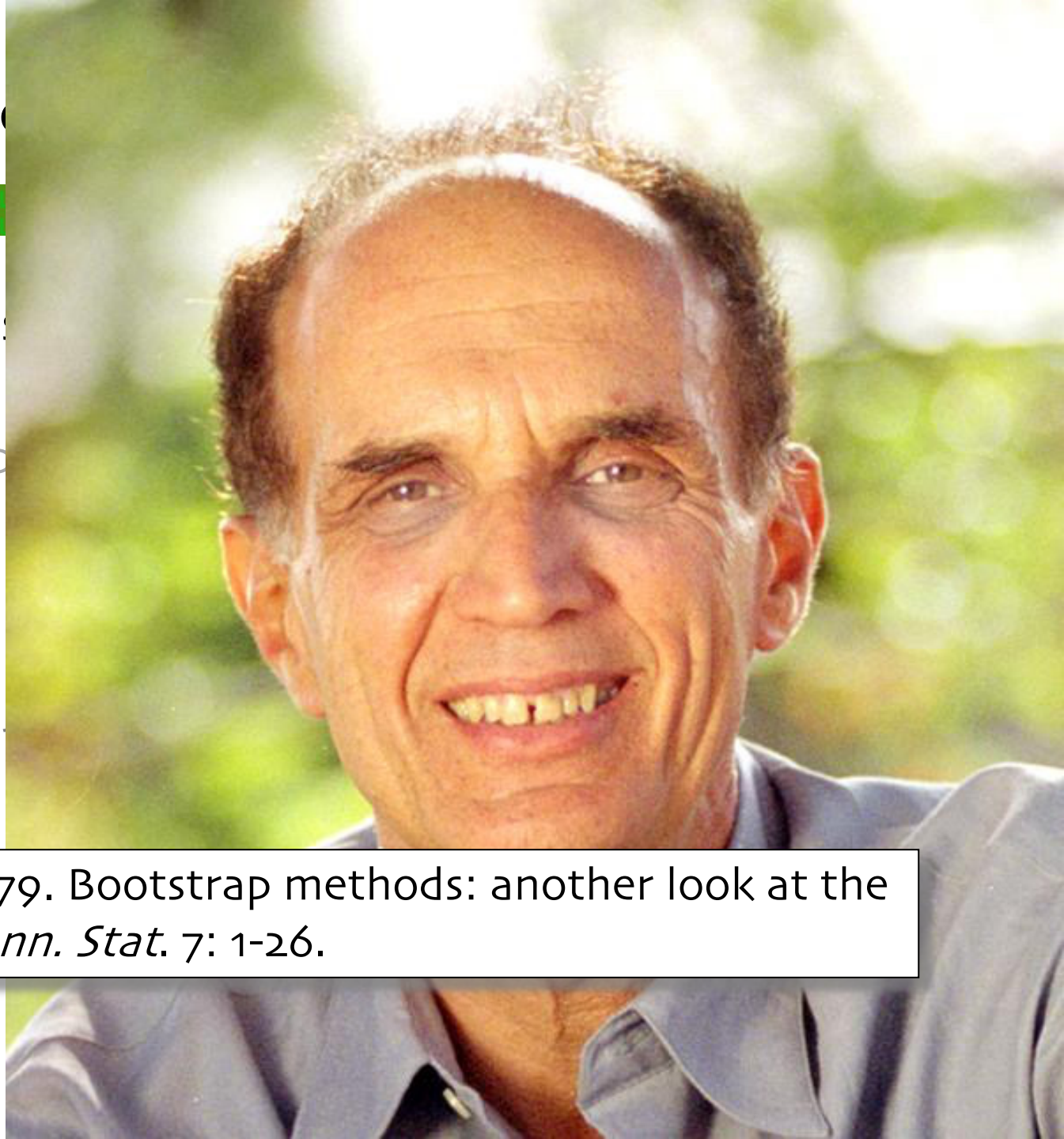
Bootstrap

Parsimony jackknifing
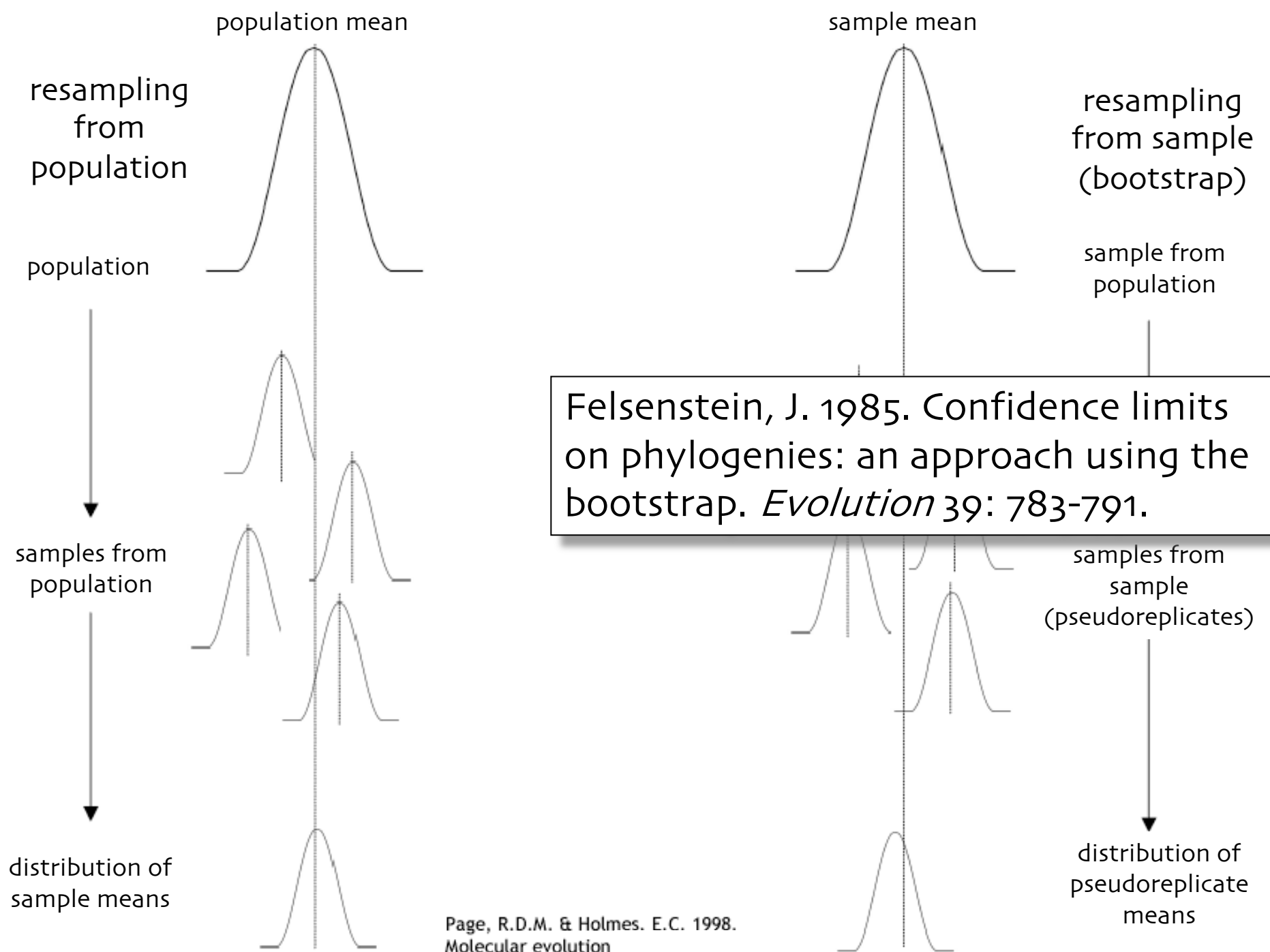
Evaluating

3 commonly us

Bremer sup

Bootstrap

Parsimony

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7: 1-26.

resampling from population

population mean

resampling from sample (bootstrap)

sample mean

population

sample from population

samples from population

samples from sample (pseudoreplicates)

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791.

distribution of sample means

distribution of pseudoreplicate means

Page, R.D.M. & Holmes. E.C. 1998. Molecular evolution

characters

	0	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	1	0	0	1	0
B	1	1	0	0	0	0	0	0	0	0
C	1	1	1	1	1	1	1	2	0	1
D	1	1	1	1	0	1	1	2	0	1
E	1	1	0	0	1	0	1	1	1	0
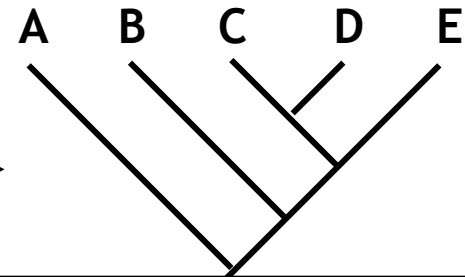
taxa

new matrix is made, equal to the size of the original one

sampling with replacement (part of the original characters will be sampled repeatedly, part will remain unsampled!)

characters

	7	8	3	7	4	5	3	0	1	9
A	0	0	0	0	0	1	0	0	1	0
B	0	1	0	0	0	0	0	0	0	0
C	2	1	1	1	1	1	1	2	0	1
D	2	1	1	1	0	1	1	2	0	1
E	1	1	0	0	1	0	1	1	1	0

taxa
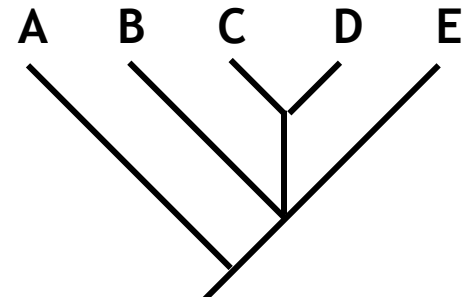
repeated several times (100- 10 000 x)

results combined to a majority rule compromise tree

disadvantage: autapomorphies, invariable characters affect the values

BOOTSTRAP

97%

52%

87%

Kitching, I.J. ym. 1998. Cladistics. 2. Painos. 228 s. Oxford University Press
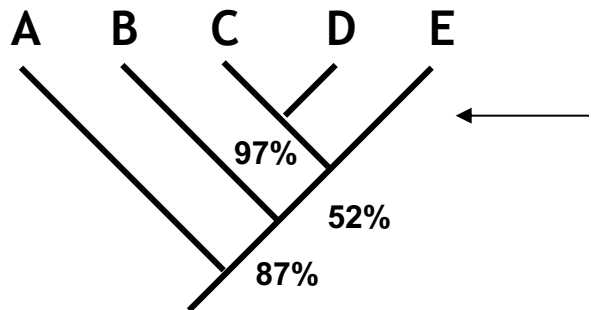
Evaluating results

3 commonly used methods:

Bremer support value

Bootstrap

Freudenstein, J.V. & Davis, J.I. 2010. Branch support via resampling:
   an empirical study. *Cladistics* 26: 643-656.

Goloboff, P.A. & Simmons, M.P. 2014. Bias in tree searches
   and its consequences for measuring group supports.
   Systematic Biology 63: 851-861.

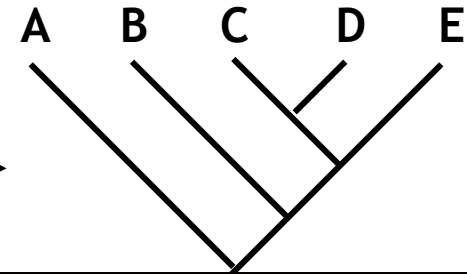# Evaluating results

3 commonly used methods:

Bremer (branch) support

Quenoille, M.H. 1949. Approximate tests of correlation in time-series. *J. R. Statist. Soc.* B 11: 68-84.

Parsimony jackknifing

Farris, J.S. & al. 1996. Parsimony jackknifing outperforms neighbor-joining. *Cladistics* 12: 99-124.

characters

	0	1	2	3	4	5	6	7	8	9
taxa A	0	0	0	0	0	1	0	0	1	0
B	1	1	0	0	0	0	0	0	0	0
C	1	1	1	1	1	1	1	2	0	1
D	1	1	1	1	0	1	1	2	0	1
E	1	1	0	0	1	0	1	1	1	0

A B C D E

new matrix is made but only PART of the original characters are sampled (sampling WITHOUT replacement)

sampling is made so that 1/e (~37%)

of characters will be deleted

characters

	0	1	3	4	7	8	9
taxa A	0	0	0	0	0	1	0
B	1	1	0	0	0	0	0
C	1	1	1	1	2	0	1
D	1	1	1	0	2	0	1
E	1	1	0	1	1	1	0
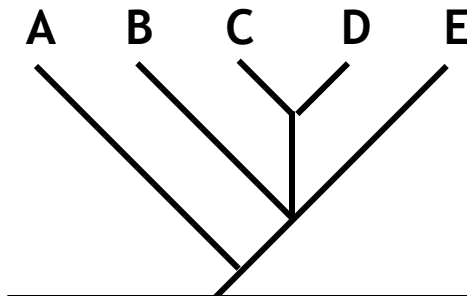
A B C D E

repeated several times (100- 10 000 x)

results combined to a majority rule compromise tree

A B C D E

97%

52%

87%

PARSIMONY JACKKNIFING

8.xi.
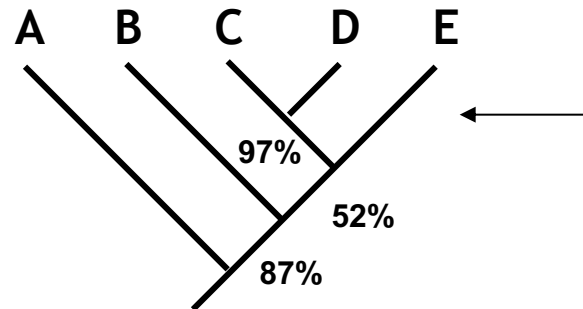
Emil Hans WILLI HENNIG

*20.4.1913 †5.11.1976

Hennig, W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik

Hennig, W. 1966. Phylogenetic systematics

CLADISTIC revolution

CLEAR, EXPLICIT & LOGICAL presentation of basic principles of phylogenetic analysis

SYNAPOMORPHY

MONOPHYLY

PARAPHYLY

TAXONOMY
systematics

description

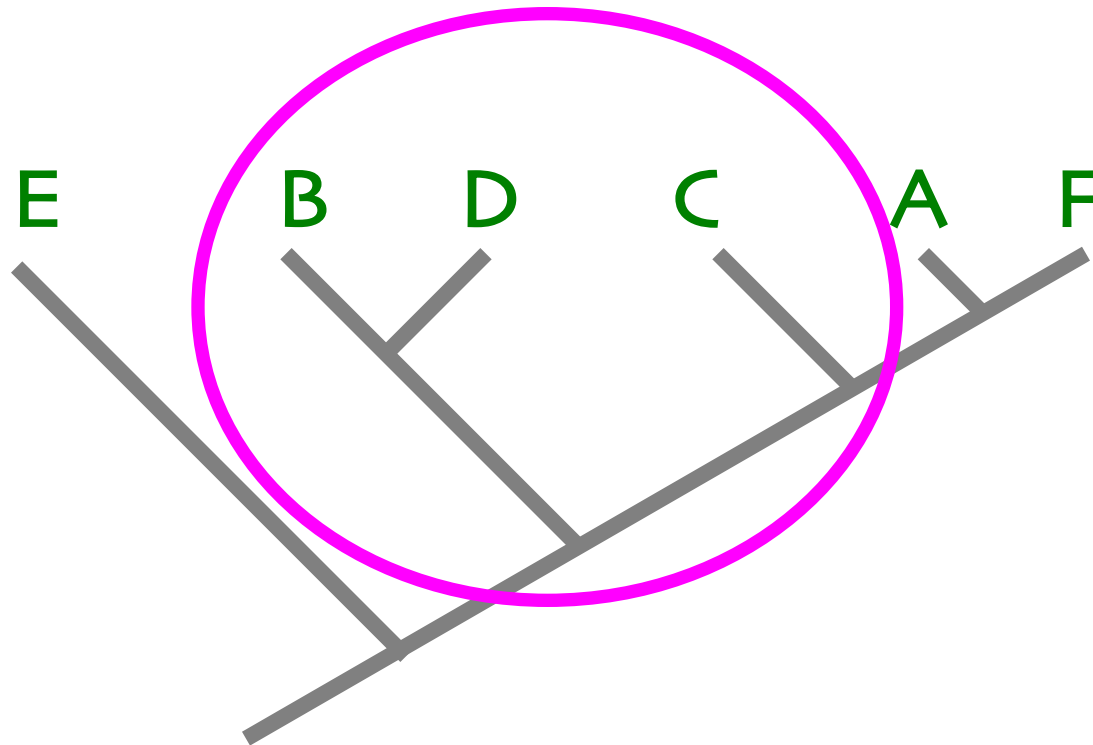nomenclature

CLASSIFICATION of organisms

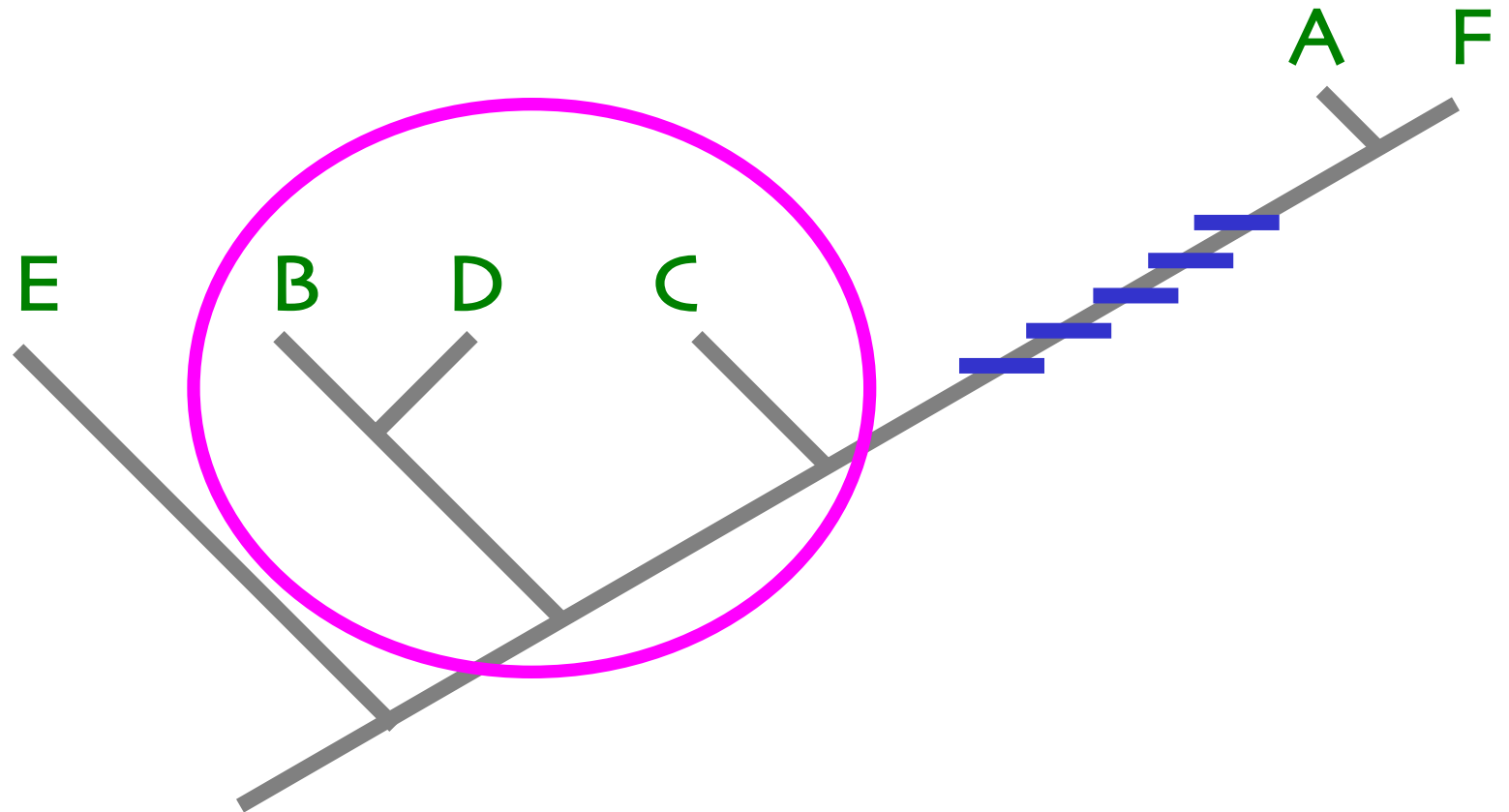enables our navigation in the ocean of biodiversity

MONOPHYLY, paraphyly, polyphyly

1. Monophyletic group includes ALL descendants of the common ancestor

2. In paraphyletic group 1 or more of the descendants are left out

MONOPHYLY, paraphyly, polyphyly

MONOPHYLY, paraphyly, polyphyly

MONOPHYLY, paraphyly, polyphyly

1. Monophyletic group includes ALL descendants of the common ancestor

2. In paraphyletic group 1 or more of the descendants are left out

3. In polyphyletic group common ancestor is left out

MONOPHYLY, paraphyly, polyphyly

E   B   D   C   A   F
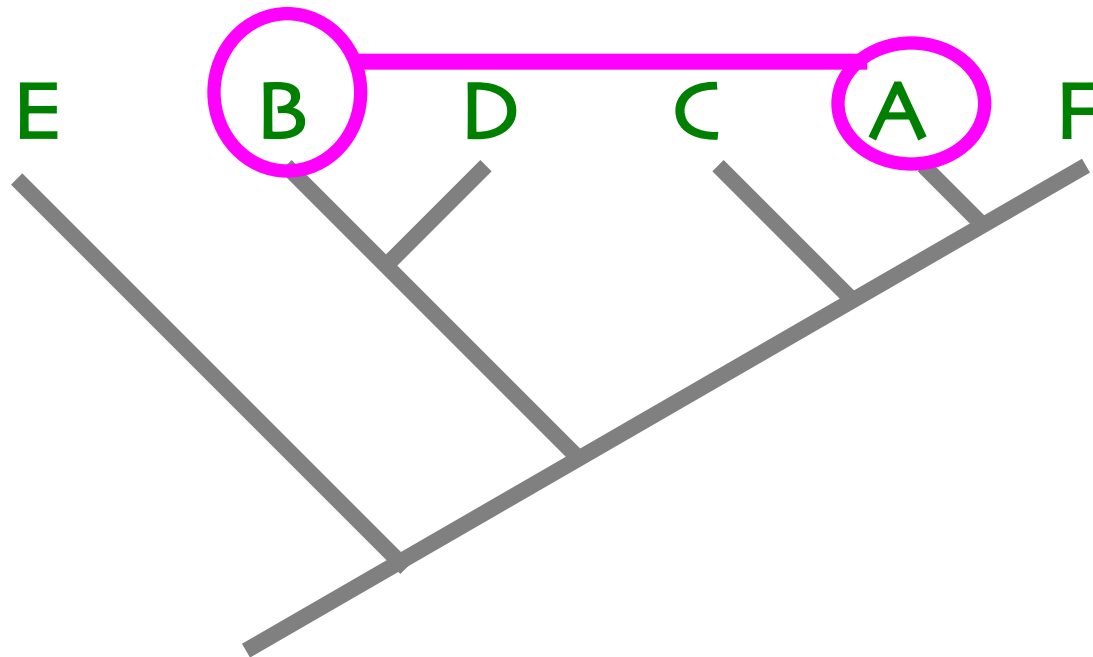
MONOPHYLY, paraphyly, polyphyly

1. Monophyletic group is defined by
   SYNAPOMORPHY

2. Paraphyletic group by **plesiomorphy**

   &

3. Polyphyletic group by **homoplasy**

MONOPHYLY, paraphyly, polyphyly

1. Monophyletic groups provide PRECISE information about relationships

2. Paraphyletic group imprecise

&

3. Polyphyletic groups **MISLEADING** information

MONOPHYLY, paraphyly, polyphyly

1. Monophyletic group can be separate from tree with 1 cut

WHOLE parts of Tree of Life

MONOPHYLY, paraphyly, polyphyly
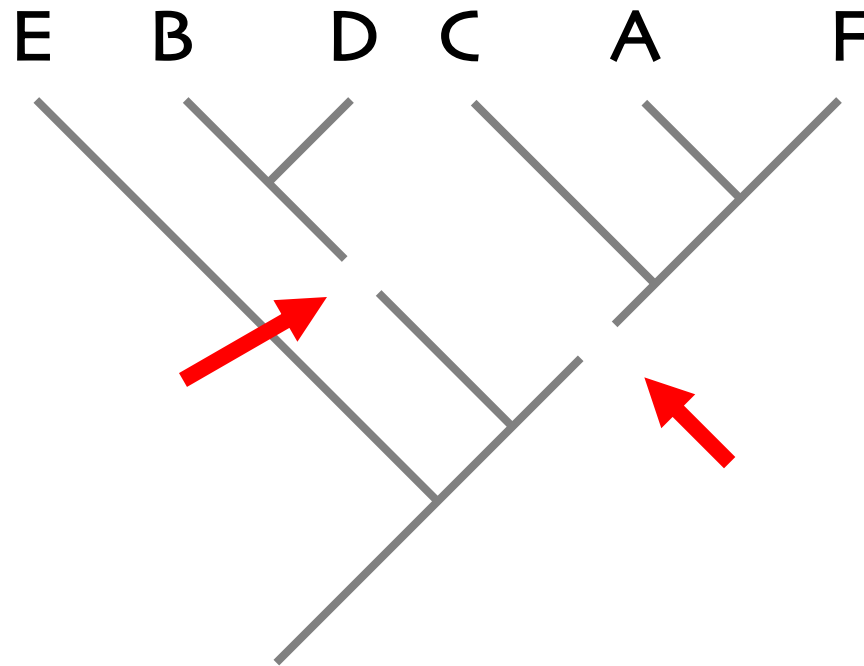
MONOPHYLY, paraphyly, polyphyly

1. Monophyletic group can be separate from tree with 1 cut

2. Paraphyletic with 2

MONOPHYLY, paraphyly, polyphyly
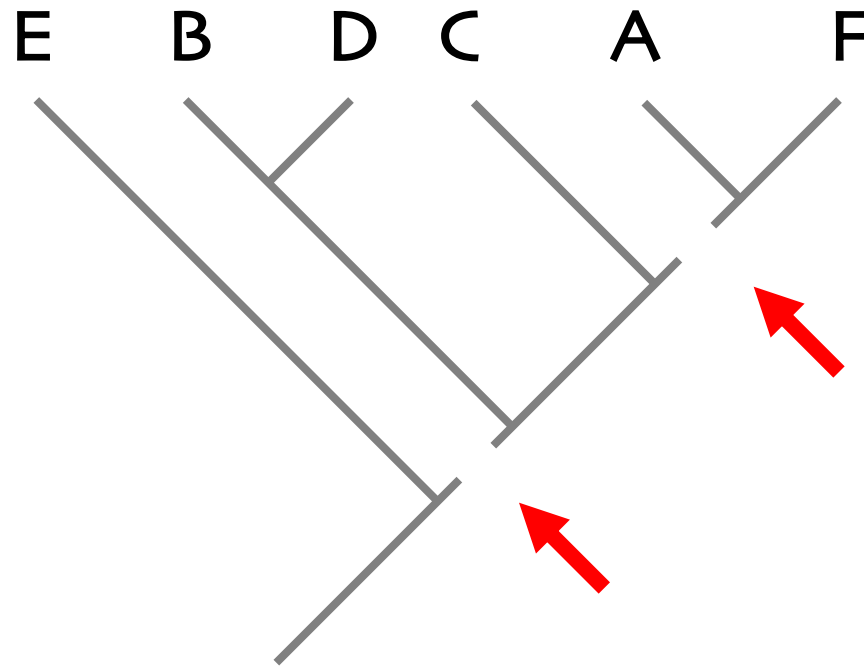
MONOPHYLY, paraphyly, polyphyly

1. Monophyletic group can be separate from tree with 1 cut

2. Paraphyletic with 2

   &

3. Polyphyletic with $\geq$ 2

MONOPHYLY, paraphyly, polyphyly
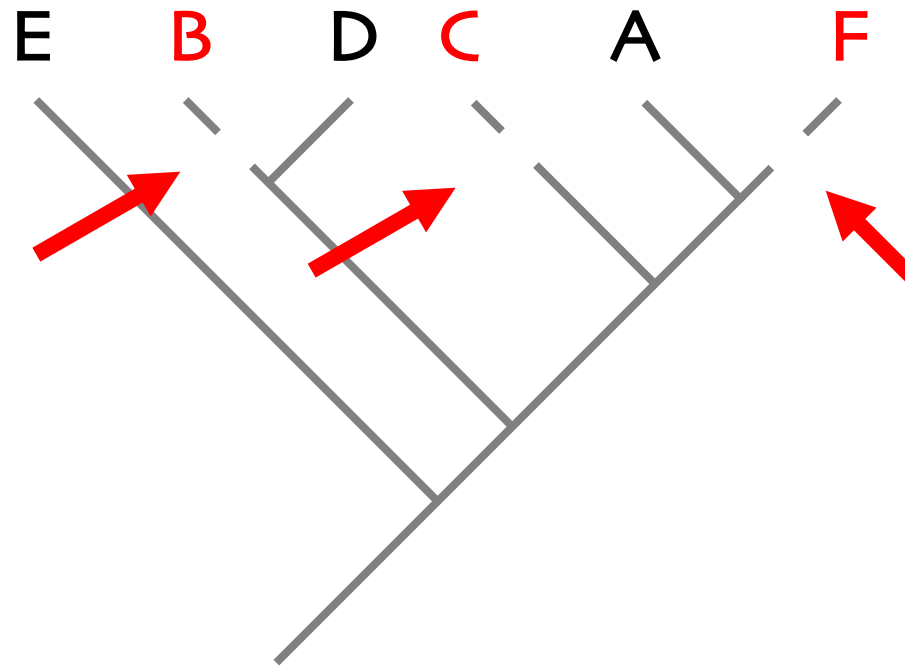
MONOPHYLY, paraphyly, polyphyly

Farris, J.S. 1974. Formal definitions of paraphyly and polyphyly. *Systematic Zoology* 23: 548-554.

defined groups inspected on the tree currently accepted as the best hypothesis about phylogeny

group membership character

1	1	0	0	0	0	0	0	0
A	B	C	D	E	F	G	H	I

member of group
ch. state = 1

does NOT belong to
group = 0

optimization (down & up)

root always signed 0

in up-pass always
assigned value of the
immediate ancestor
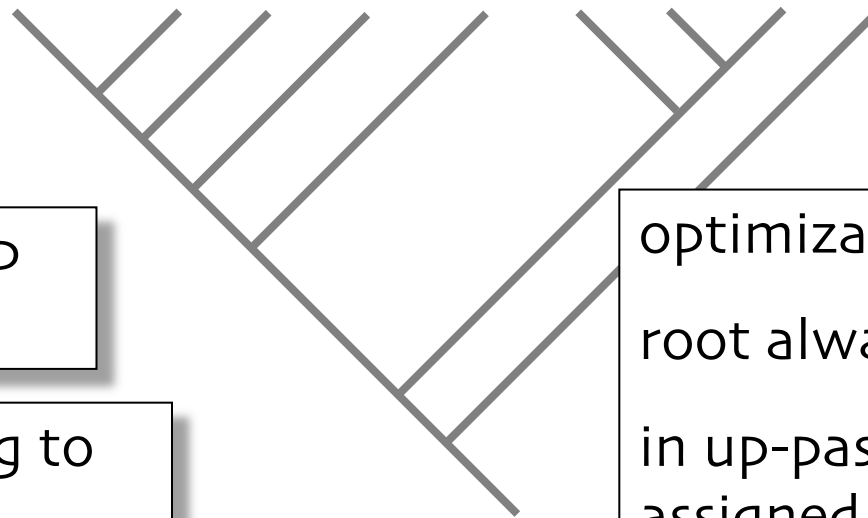
MONOPHYLY, paraphyly, polyphyly

change o > 1 only once -->

group monophyletic (A & B)
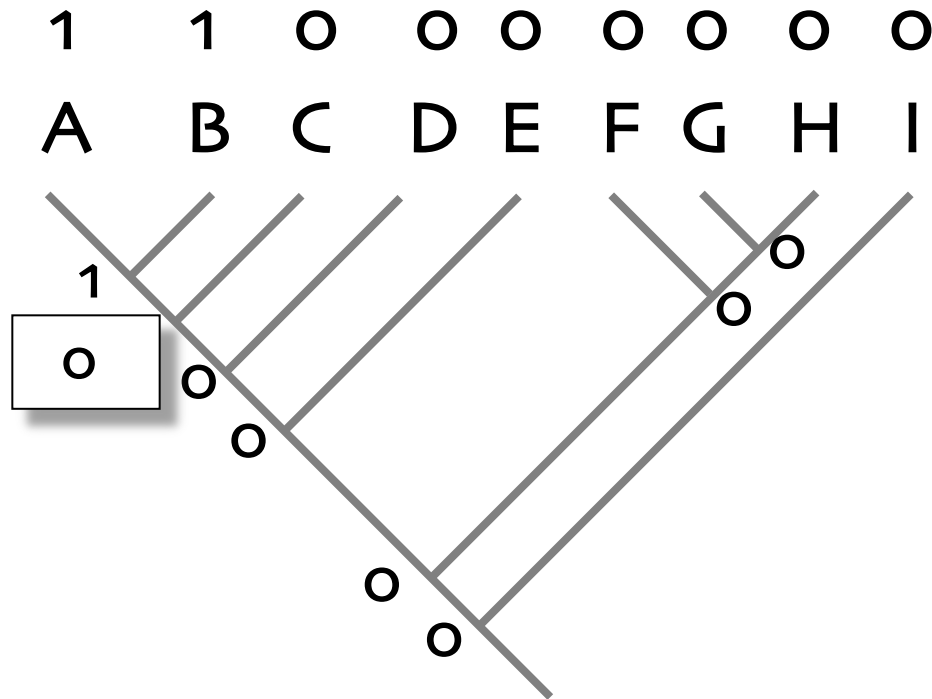
MONOPHYLY, paraphyly, polyphyly

change 1 > o   -->

group paraphyletic (C, D & E)

o   o   1   1   1   o   o   o   o

A   B   C   D   E   F   G   H   I

o

1

1

1

o

o

o

o
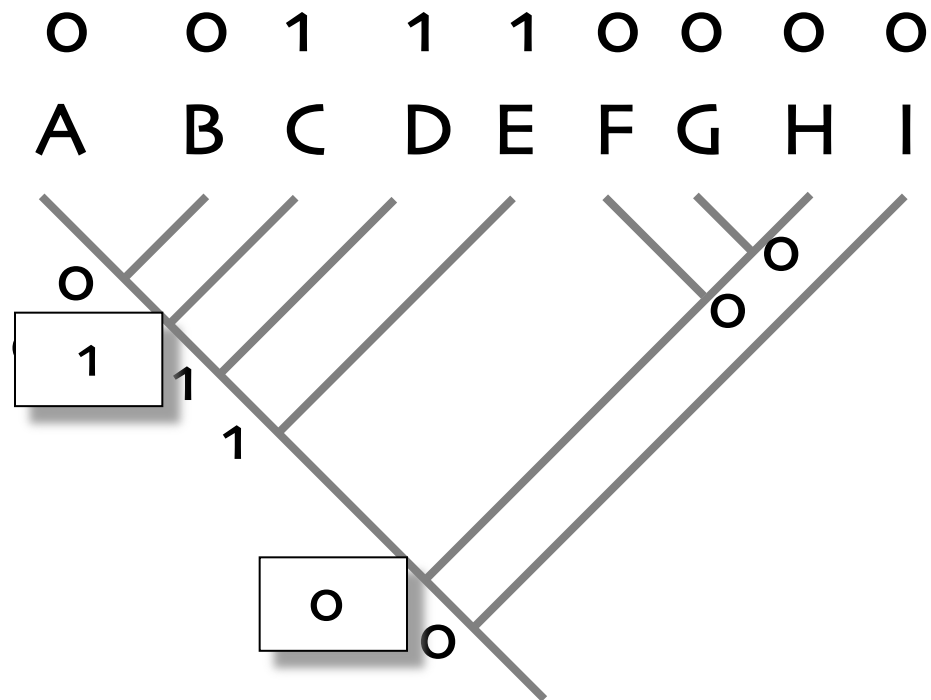
MONOPHYLY, paraphyly, polyphyly

group membership matrices
can be produced with e.g. TNT

all other changes -->

group polyphyletic (D & F)

0   0   0   1   0   1   0   0   0

A   B   C   D   E   F   G   H   I
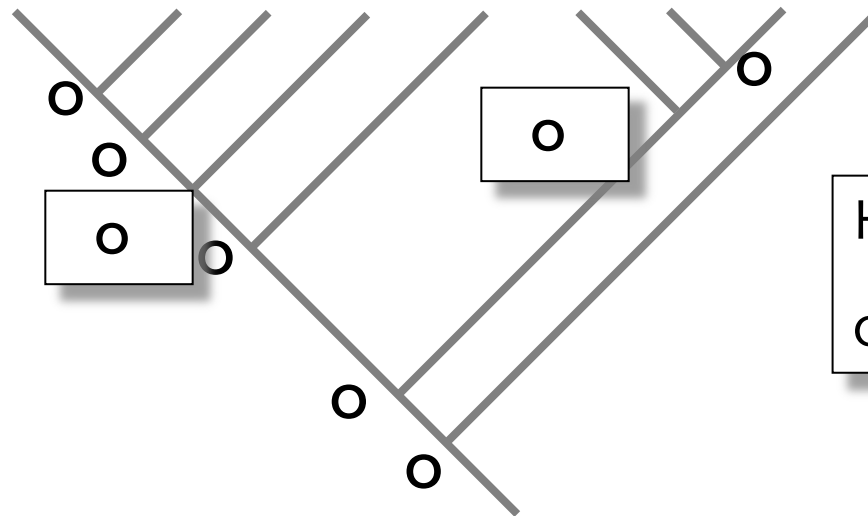
Here change

0 > 1   2X

8.xi.

Optimality criteria

1. evolutionary distance   (Un)Weighted Pair-Group Method using Arithmetic averages

   e.g. UPGMA, WPGMA, NJ   neighbor-joining

   advantage SPEED

   disadvantages:
   - information reduced to SINGLE numerical value

   - ultrametricity or additivity required   NEVER/rarely present in real material

   - NEGATIVE branch lengths

Manhattan vs. Euclidean distance

$$m(p_1, p_2) = |x_1 - x_2| + |y_1 - y_2|$$

13

Ford, G. 1993. Lecture notes on engineering measurement for software engineers. Carnegie Mellon Univ.

Manhattan vs. Euclidean distance

8, 5

0, 0
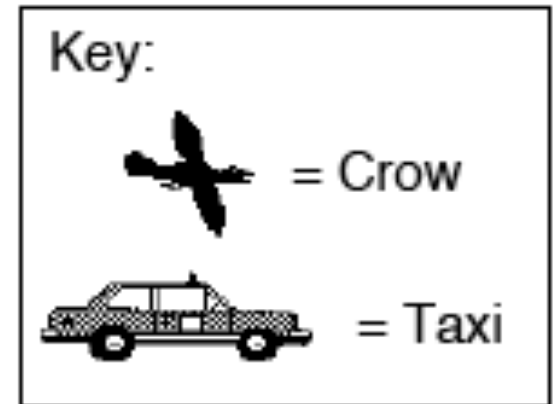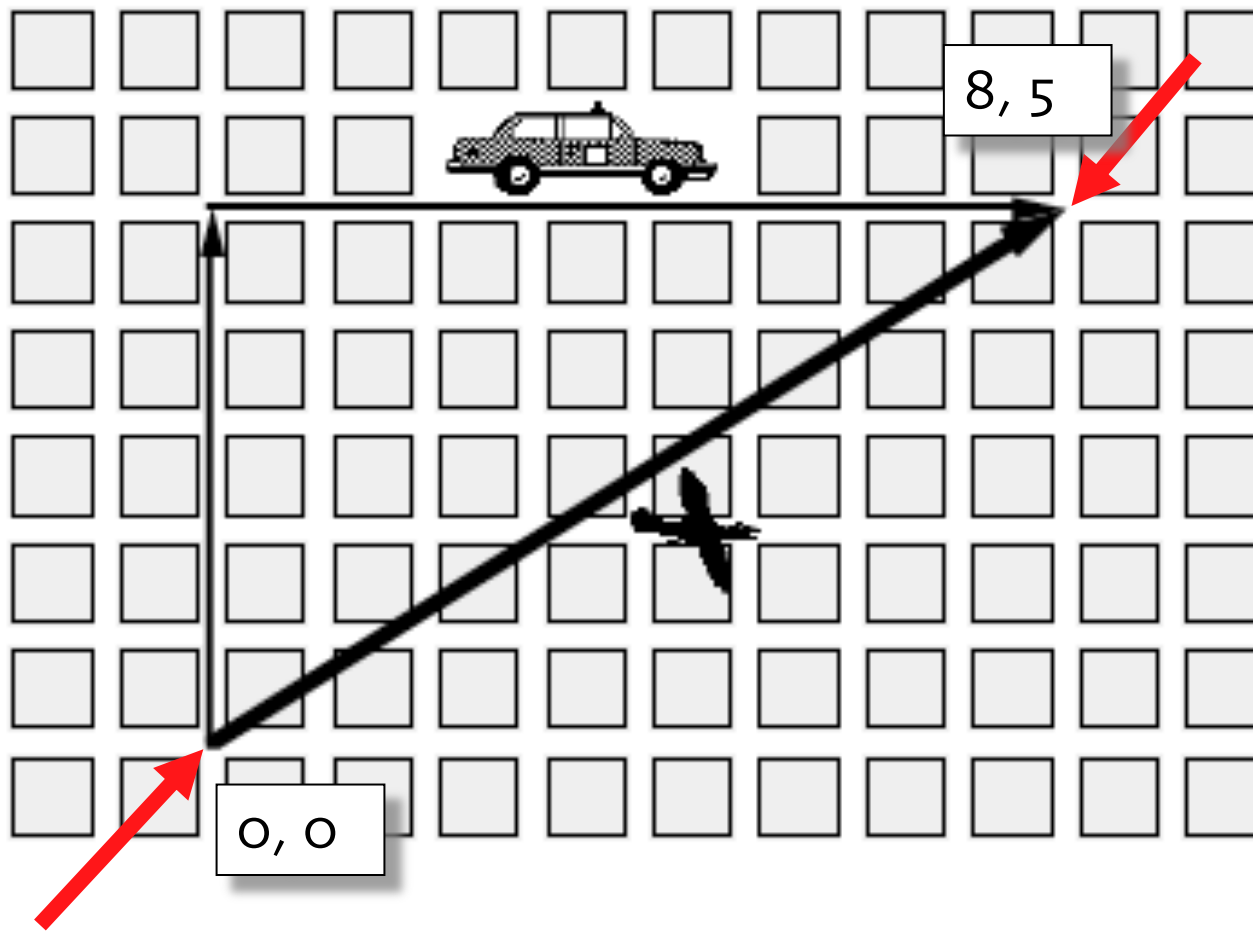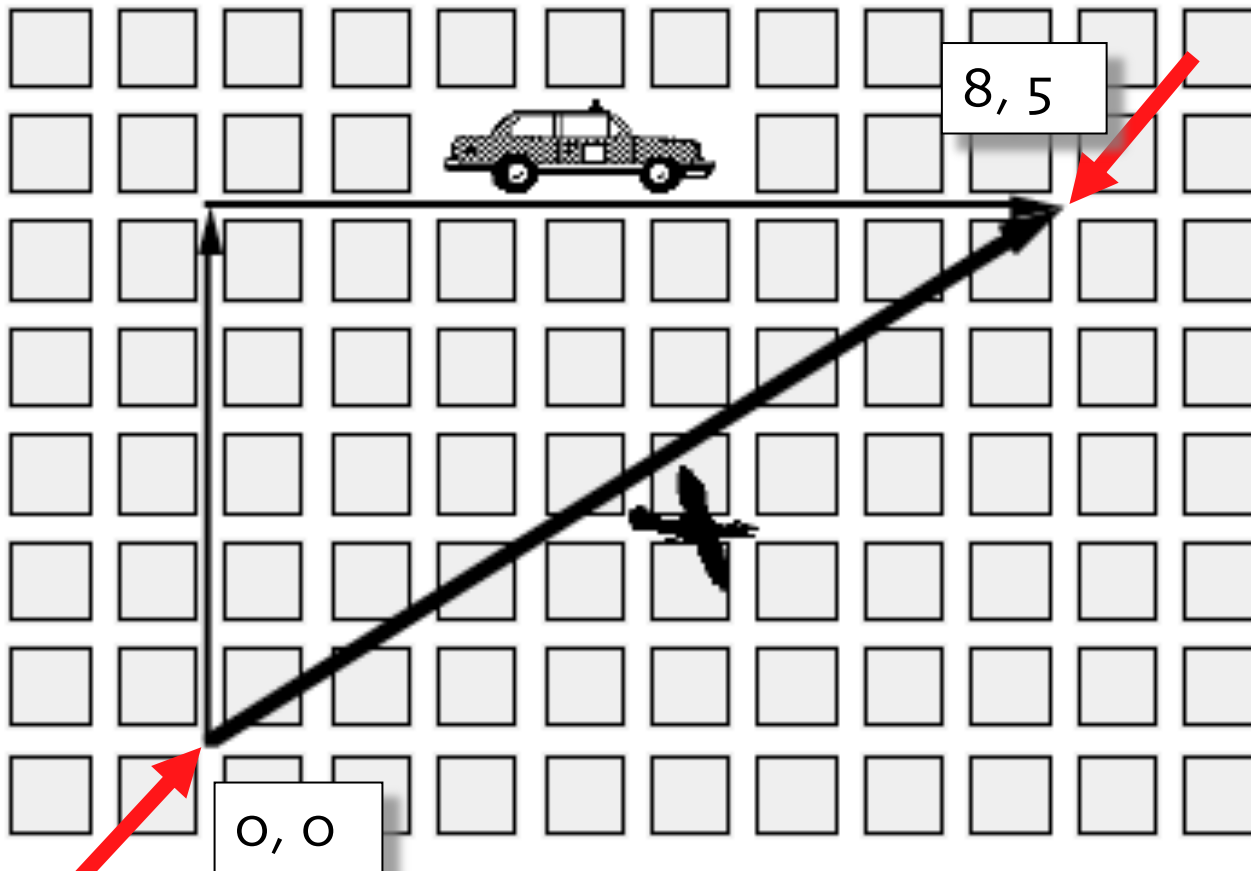
Key:

= Crow

= Taxi

$$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

9,43

	A	B	C
A	0	1	1
B		0	1
C			0

	A	B	C
A	0	1	1
B		0	1
C			0

Farris, J.S. 1985. Distance data revisited. *Cladistics* 1: 67-85.

Farris, J.S. 1986. Distances and statistics. *Cladistics* 2: 144-157.

C

G

A

C

0,5

T B

0,5

0,5

X
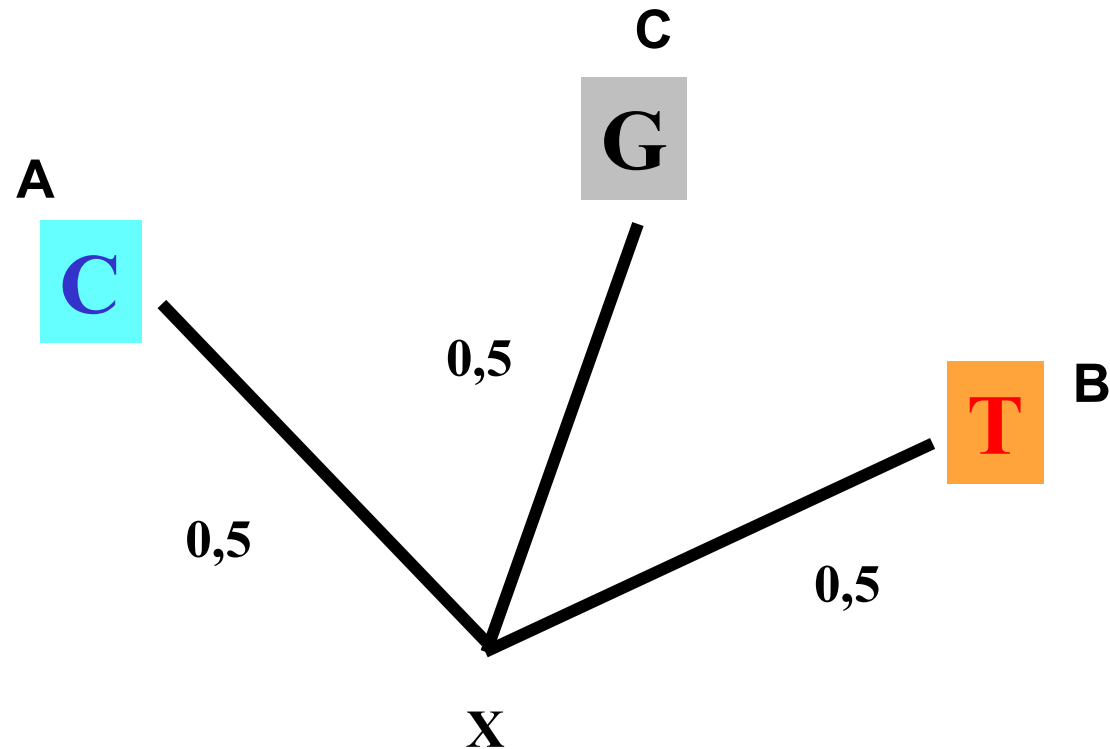
sum of branch lengths = 1,5

BUT at least **2 SUBSTITUTIONS** must have happened to produce variation observed

Optimality criteria

1. evolutionary distance

2. parsimony

comparison of hypotheses is based on optimization of individual CHARACTERS

information about individual characters RETAINED

return from tree to matrix possible

ALL terminals treated SIMULTANEOUSLY, not pair-wise

Parsimony or models?

optimality criteria:

1. evolutionary distance

2. parsimony

3. model based methods

   MODEL of evolution of explicitly defined

Parsimony or models?

Parsimony analysis

what is the best tree (= hypothesis) based on characters, background knowledge & optimality criterion

descent with modification

EXPLANATORY power of the data MAXIMIZED

model of evolution of explicitly defined

Model based analyses – what is a tree with highest likelihood
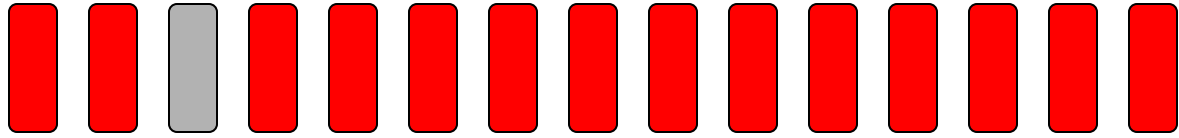
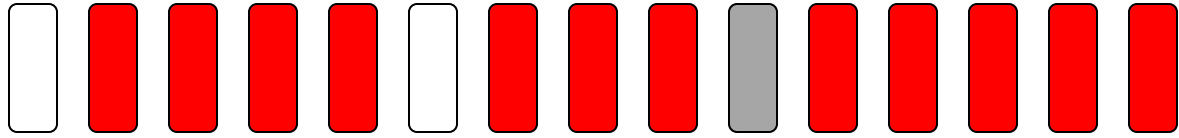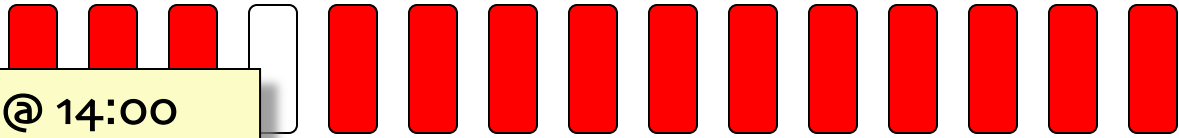Maximum likelihood

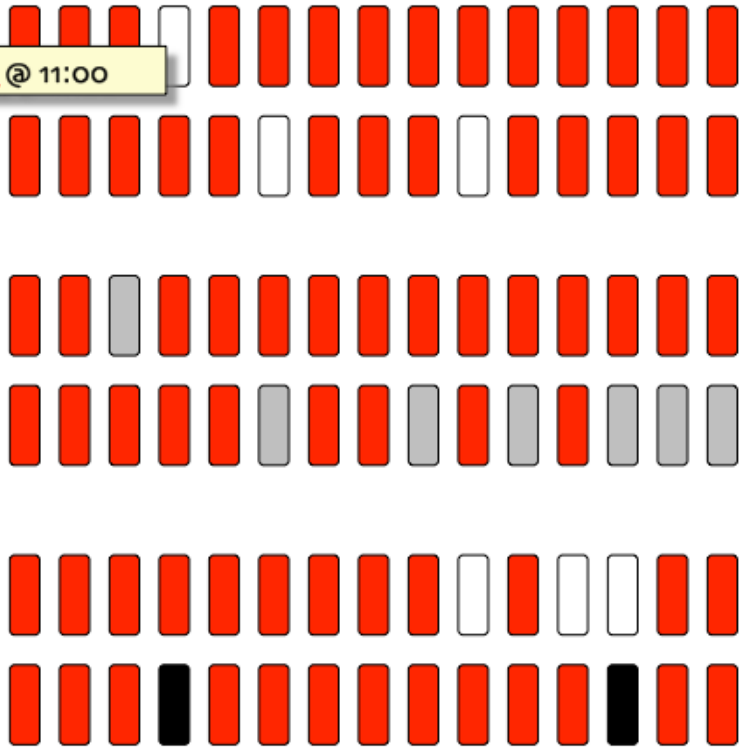models accepted *a priori*

Bayesian analyses

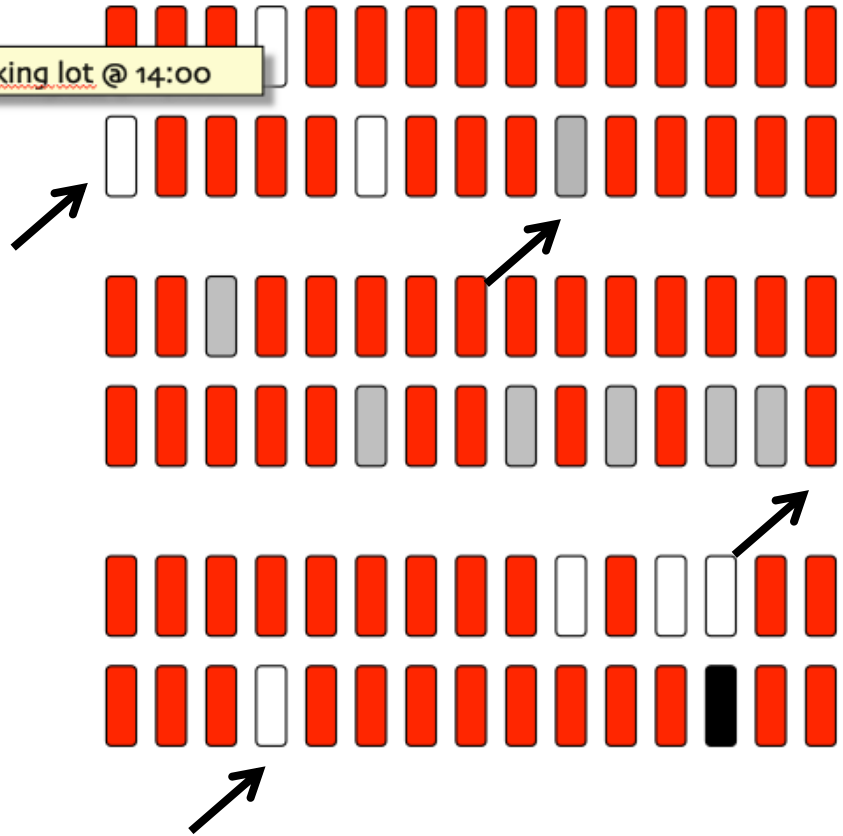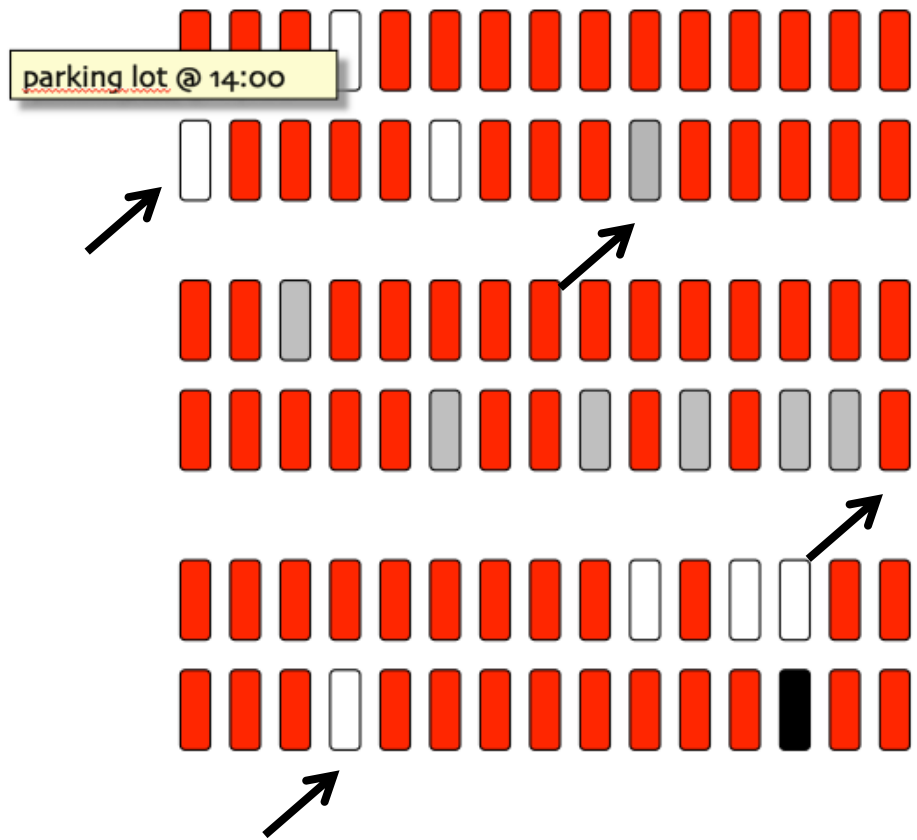WHY should models be incorporated into our analyses?
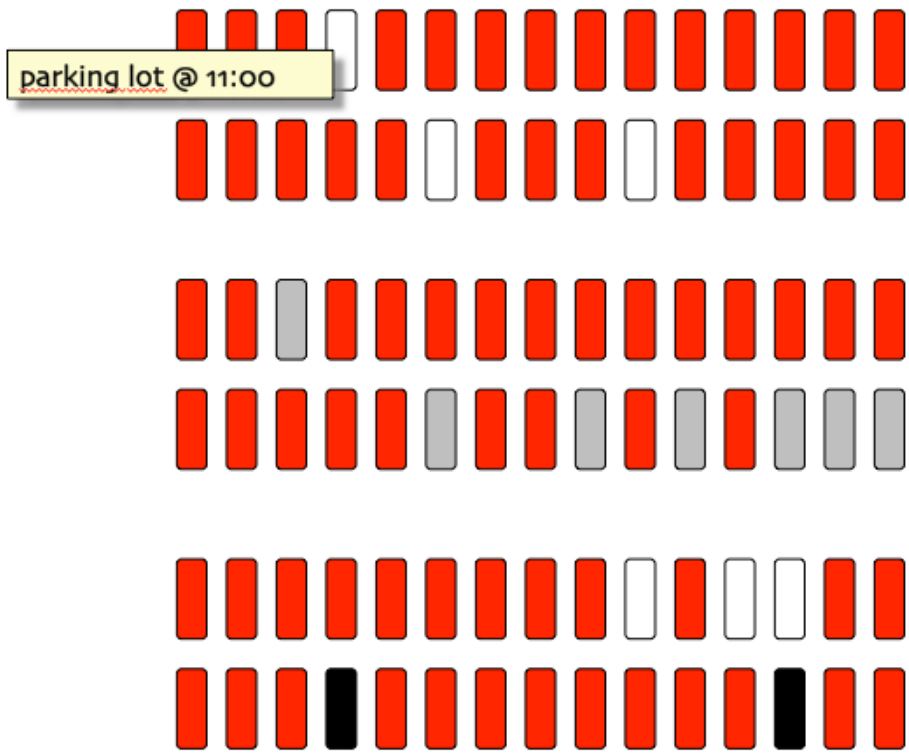
parking lot @ 11:00

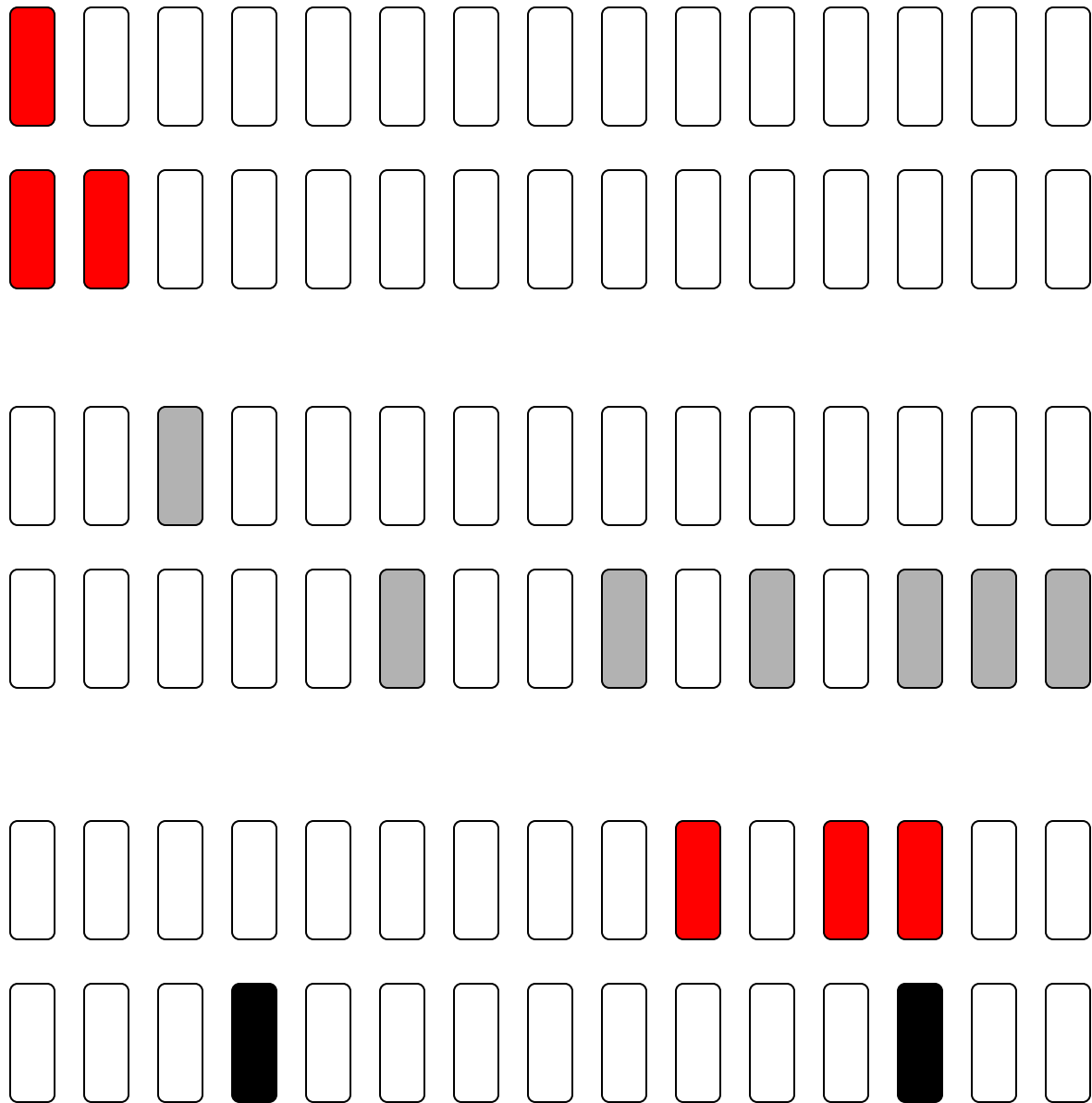parking lot @ 14:00
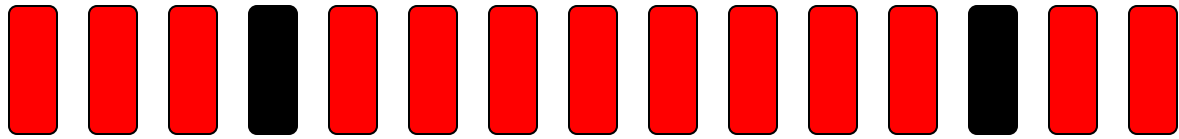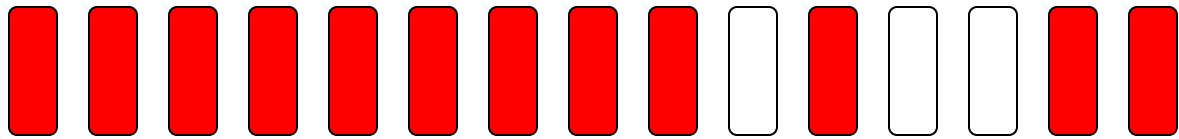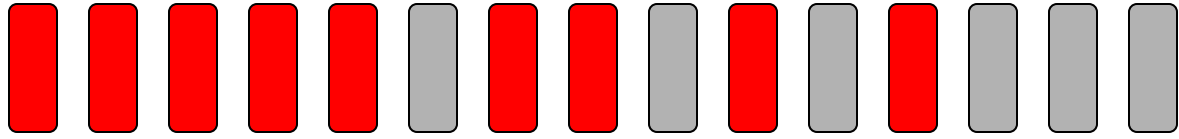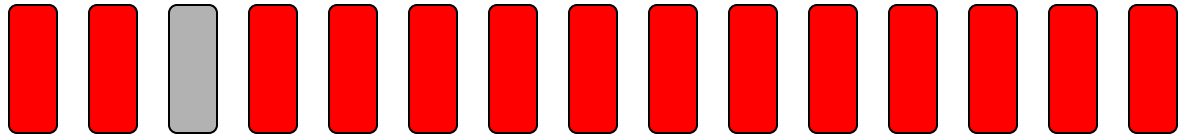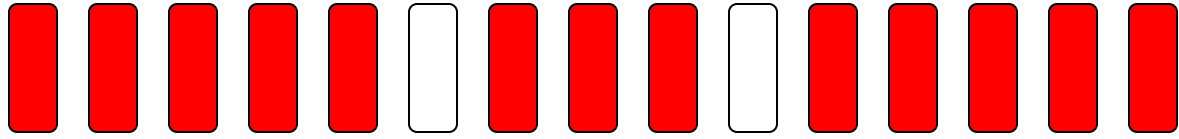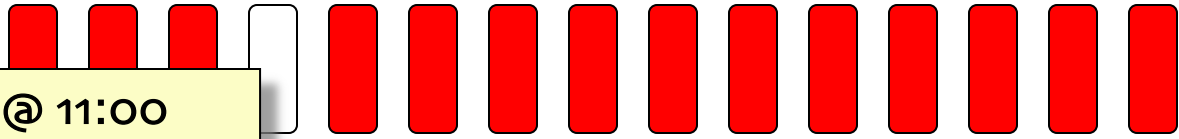
parking lot @ 11:00
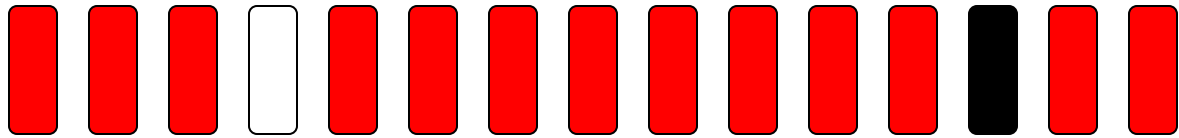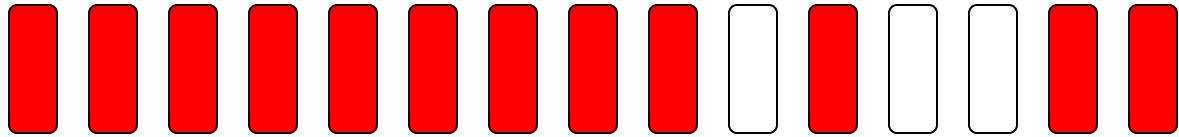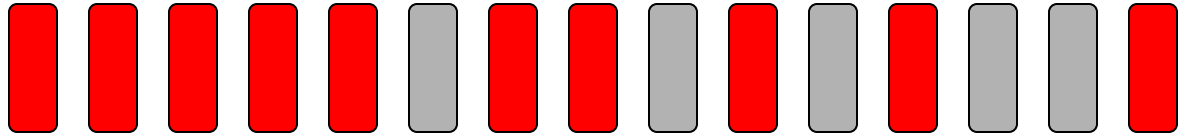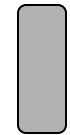
parking lot @ 14:00

parking lot @ 11:00

parking lot @ 14:00

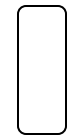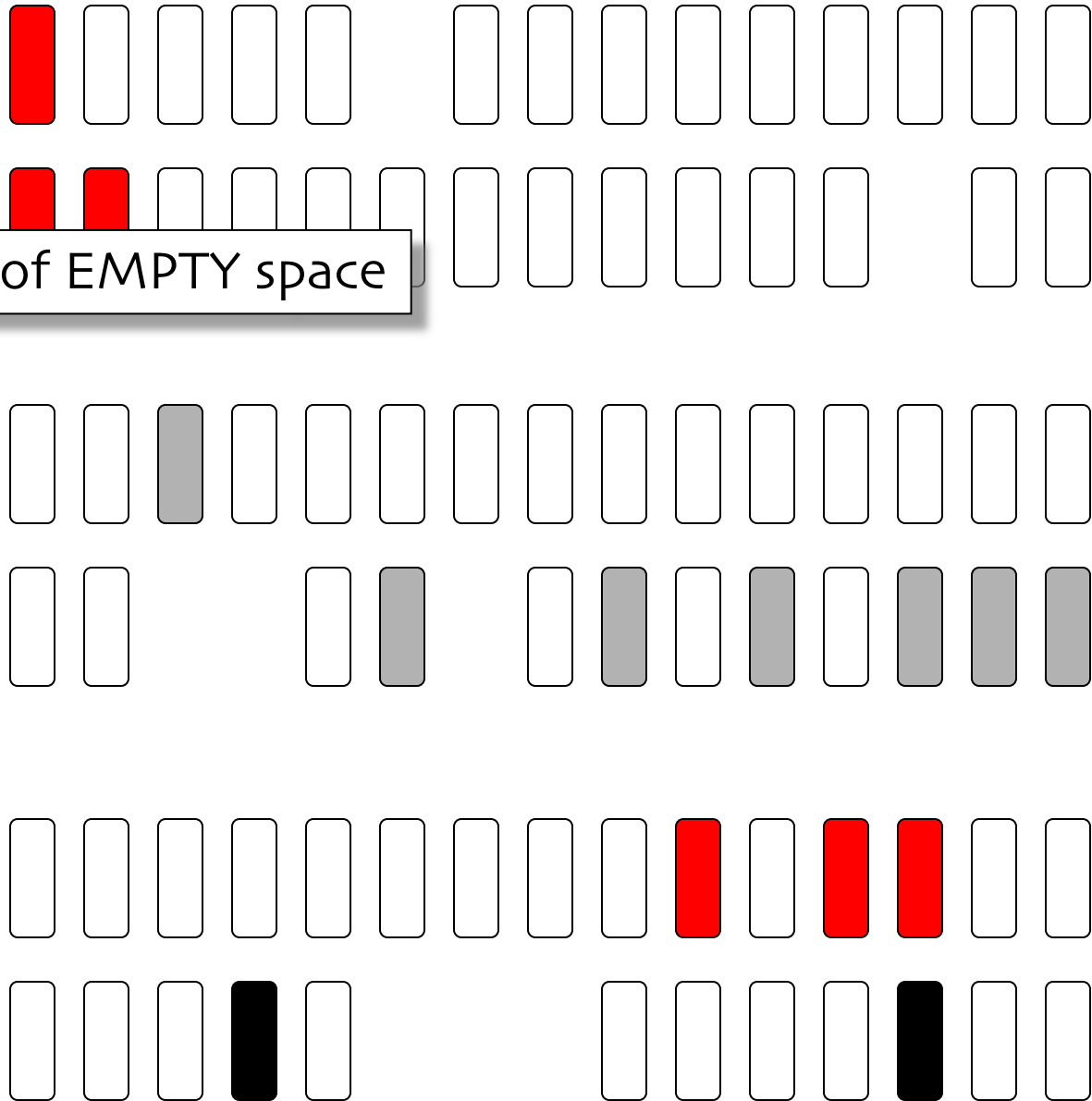6%

2%

7%

85%

likelihood of EMPTY space

6%

2%

7%

85%
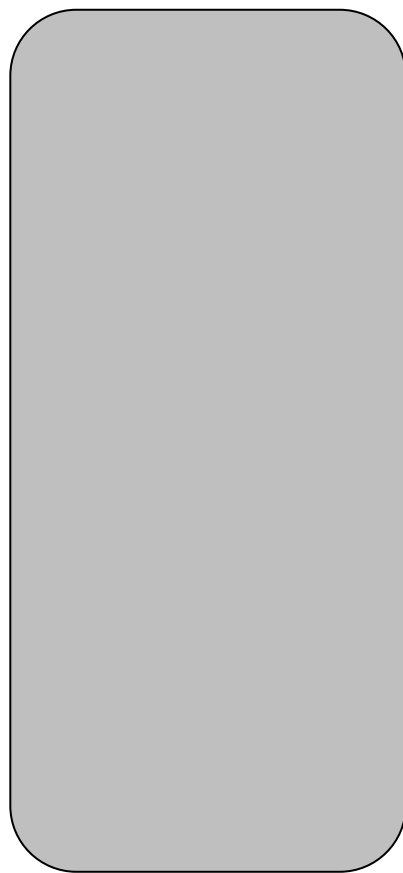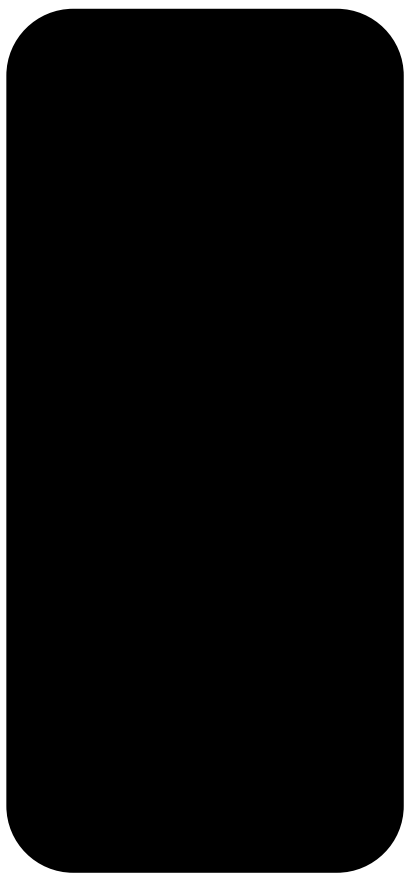
Steps in sites in tree

Calculated over tree

Site
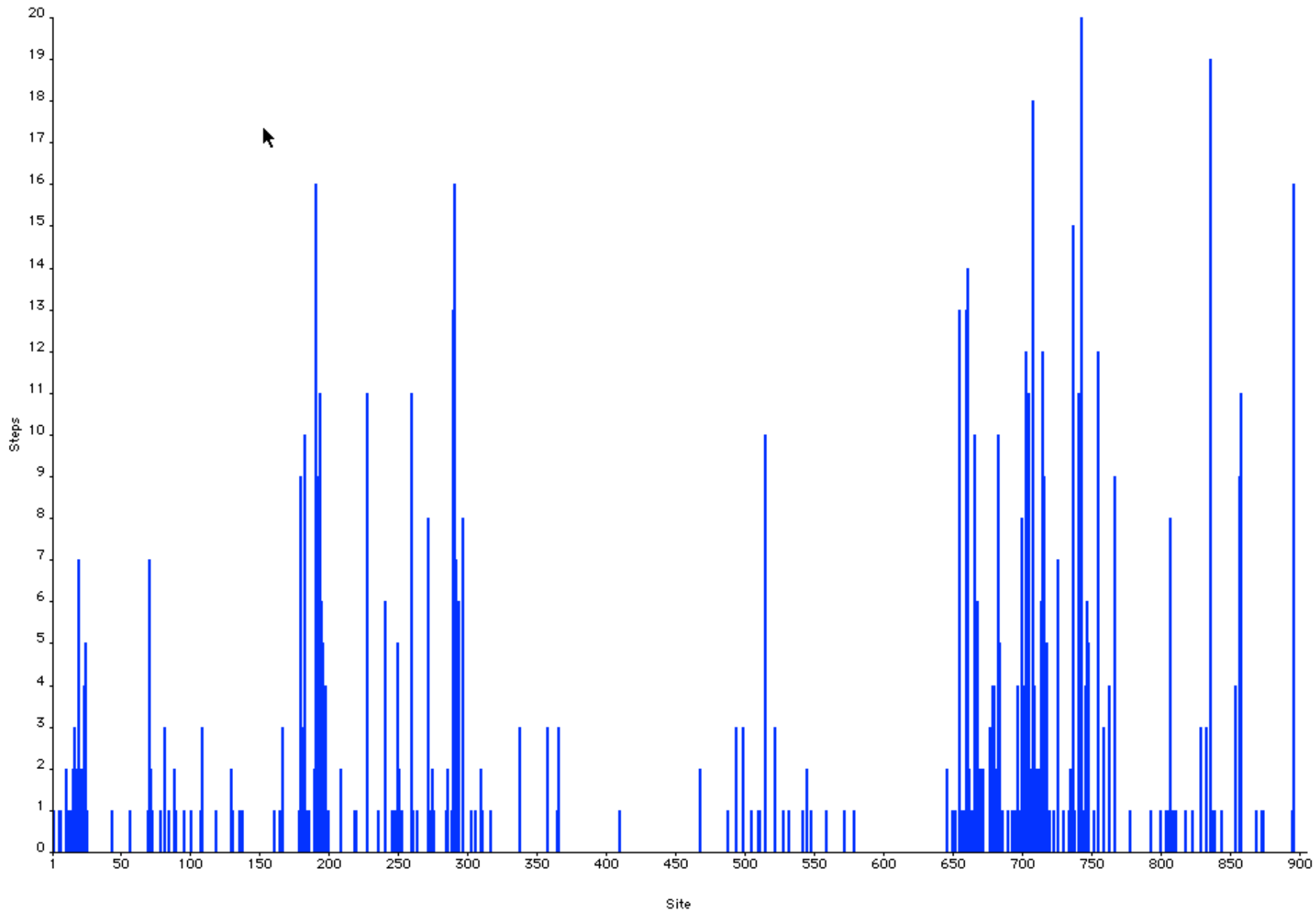
Parsimony or models?

can unique events of history be MODELLED?

in parsimony analyses emphasis is on ever more
comprehensive heuristics of **tree search**

in model based analyses emphasis is on developing
more sophisticated MODELS

analyses are VERY demanding computationally and
thus in many cases heuristics of tree search are
deliberately VERY **superficial** | e.g. RAxML |

with NCM model parsimony and ML trees are identical

| no common mechanism |

| all characters with same
number of states |

Parsimony or models?

Efron, B. 2013. Bayes theorem in the 21st century. *Science* 340: 1177-1178.

Tuffley, C. & Steel, M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.* 59: 581-607.

Goloboff, P.A. 2003. Parsimony, likelihood, and simplicity. *Cladistics* 19: 91-103.

Simmons, M.P. 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. *Mol. Phyl. & Evol.* 80: 267-80.

Simmons M.P. & Goloboff, P.A. 2014. Dubious resolution and support from published sparse supermatrices: the importance of thorough tree searches. *Mol. Phyl. & Evol.* 80: 334-348.

Farris, J.S. 1973. A probability model for inferring evolutionary trees. *Syst. Biol.* 22: 250-256.

SUMMARY

monophyly is one of the CENTRAL PRINCIPLES of cladistics

> ONLY MONOPHYLETIC groups (=clades) provide precise information about relationships

consensus tree is ONLY SUMMARY of numerous trees

> all trees are **NOT** equal

three commonly used indices to evaluate results

> numeric values obtained are dependent on thoroughness of search used in finding them

while ALL of these are commonly used their status and importance is still in dispute

> NO logical connection to results obtained based on analyses of REAL & ALL data

SUMMARY

at least they ARE ABLE to show parts of the tree
with the **WEAKEST** hypotheses, parts of trees
easiest to refute? candidates for more detailed study!

optimal tree(s) should <u>**NOT**</u> be changed based on
support values   still the BEST hypothesis

it is IMPOSSIBLE to know whether our estimates are
truthful   UNIQUE nature of history, testing results obtained
NOT possible

nomothetic vs idiographic sciences
generalities & laws vs contingent & unique

history includes also unlikely events applicable also to phylogeny

if something is highly unlikely it does NOT mean that it is
IMPOSSIBLE

HOME "QUIZ"

Why, in this age of whole genome sequences, it is still relevant to study & use also morphological characters in phylogenetic analyses?

Why does the cutting of continuous sequence stretches within CONSERVED lead into huge speed-up in DO analyses?

come up with arguments **for the 1st demo Mon 14.xi.**