

19.xi.



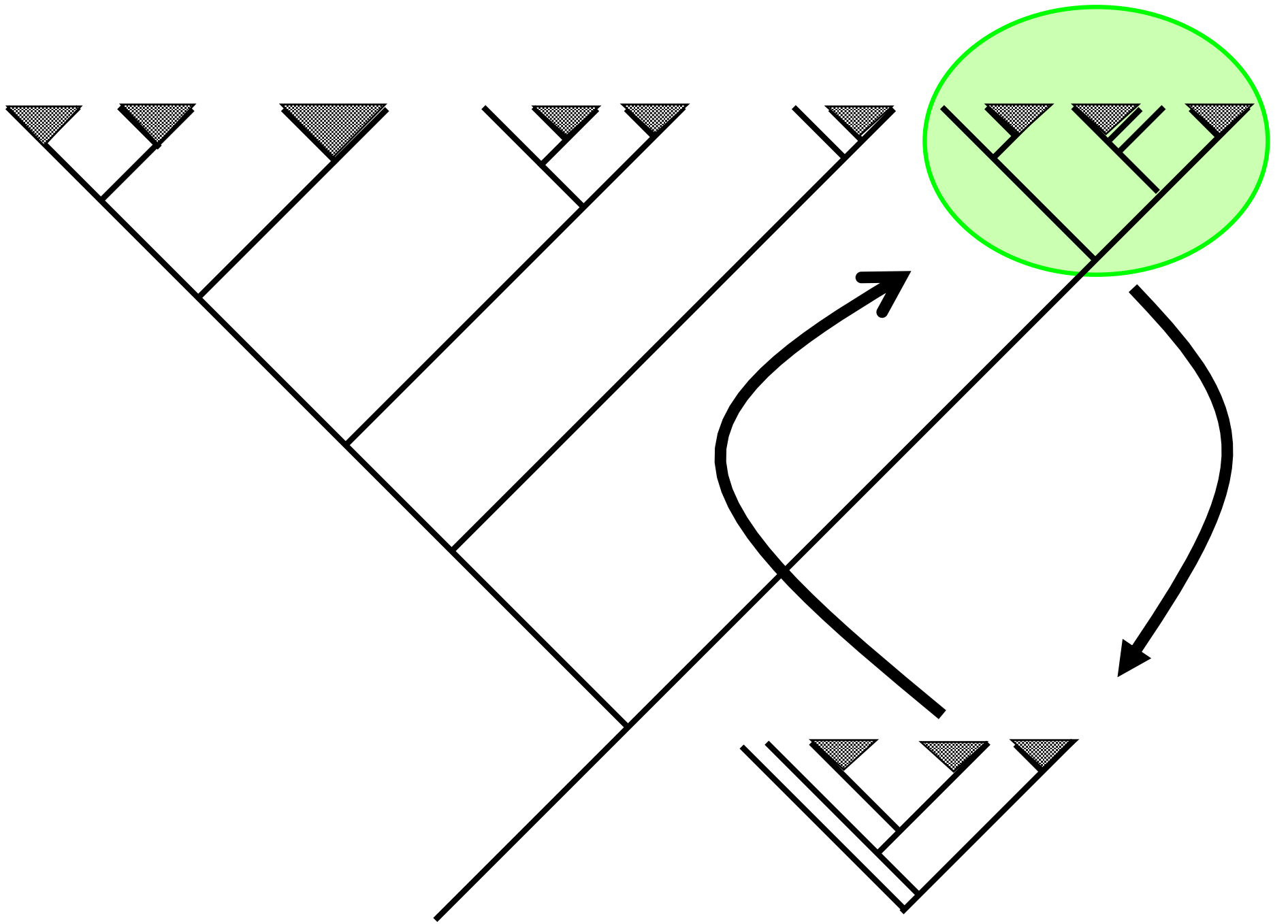
1. new search strategies
2. evaluating results
3. summary

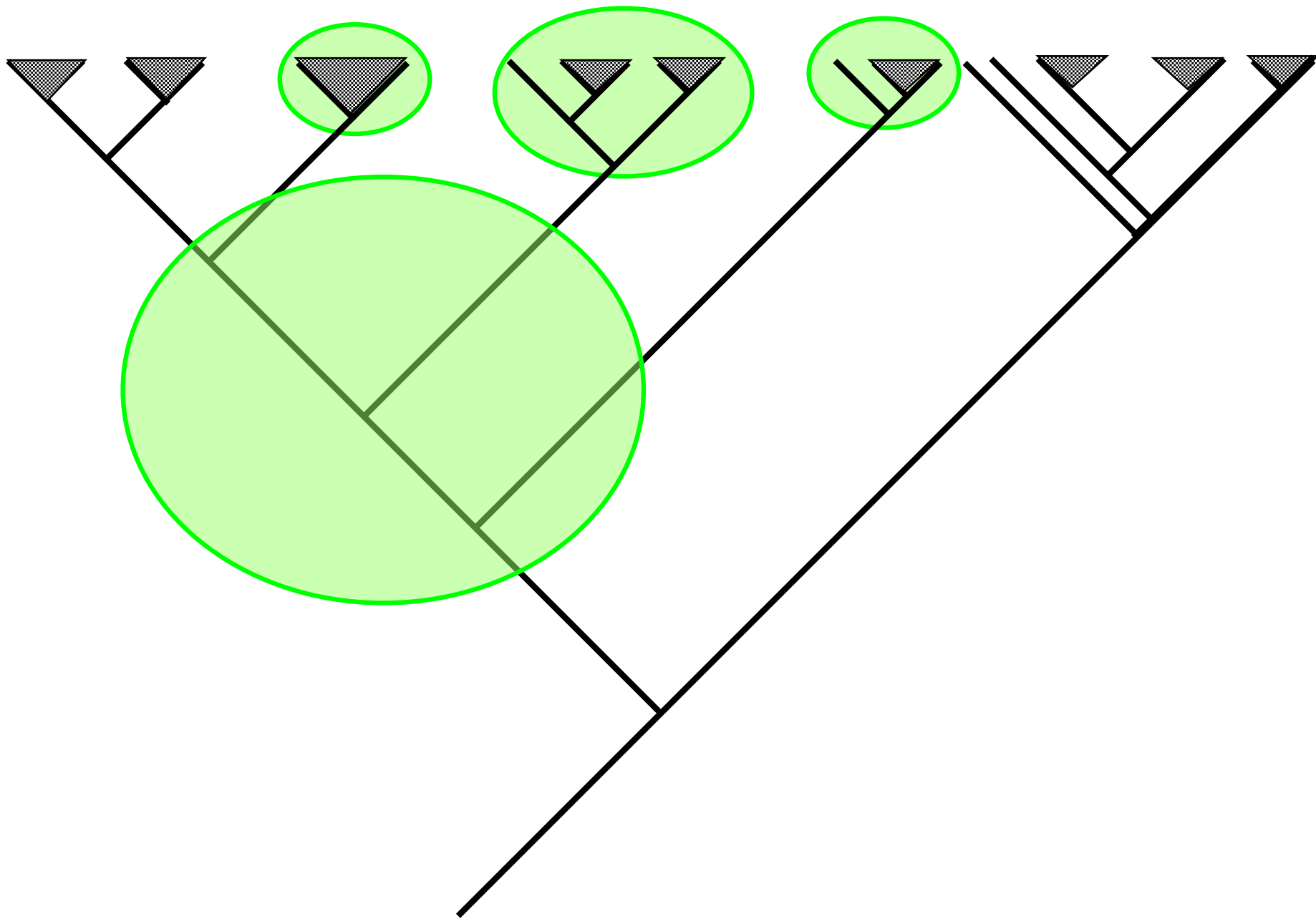
Goloboff, P. Sectorial search

1. part (a sector) of the tree found by traditional search chosen
2. a LOCAL search performed in this part & returned to original tree
3. search performed numerous times with different parts
4. leads only rarely in finding optimal solution but still MUCH faster than TBR

alternatives:	RSS (random sectorial searches)
	CSS (consensus-based sectorial s.)
	MSS (mixed sectorial s.)

processor time increases LINEARLY in relation to number of sectors EXPONENTIALLY in relation to number of terminals in TBR





Random sectorial search

1. a sector including n terminals chosen randomly
2. LOCAL analyses x times (RAS+TBR), only 1 tree saved, if shorter trees found continue to 3., otherwise add repetitions
3. shortest solution for this sector returned to whole tree
4. analysis of whole material, return to 1. & repeated certain number of times

Consensus based sectorial search (CSS)

1. based on consensus tree a sector with polytomy
reveals conflict
2. local searches (RAS+TBR) made n times with saving in cache only 1 tree, if shorter trees not found go 3. otherwise add number of searches
3. shortest tree & topology of the sector included saved, return to 1., repeat m times
4. move parts of the WHOLE tree, return to 1., repeat z times

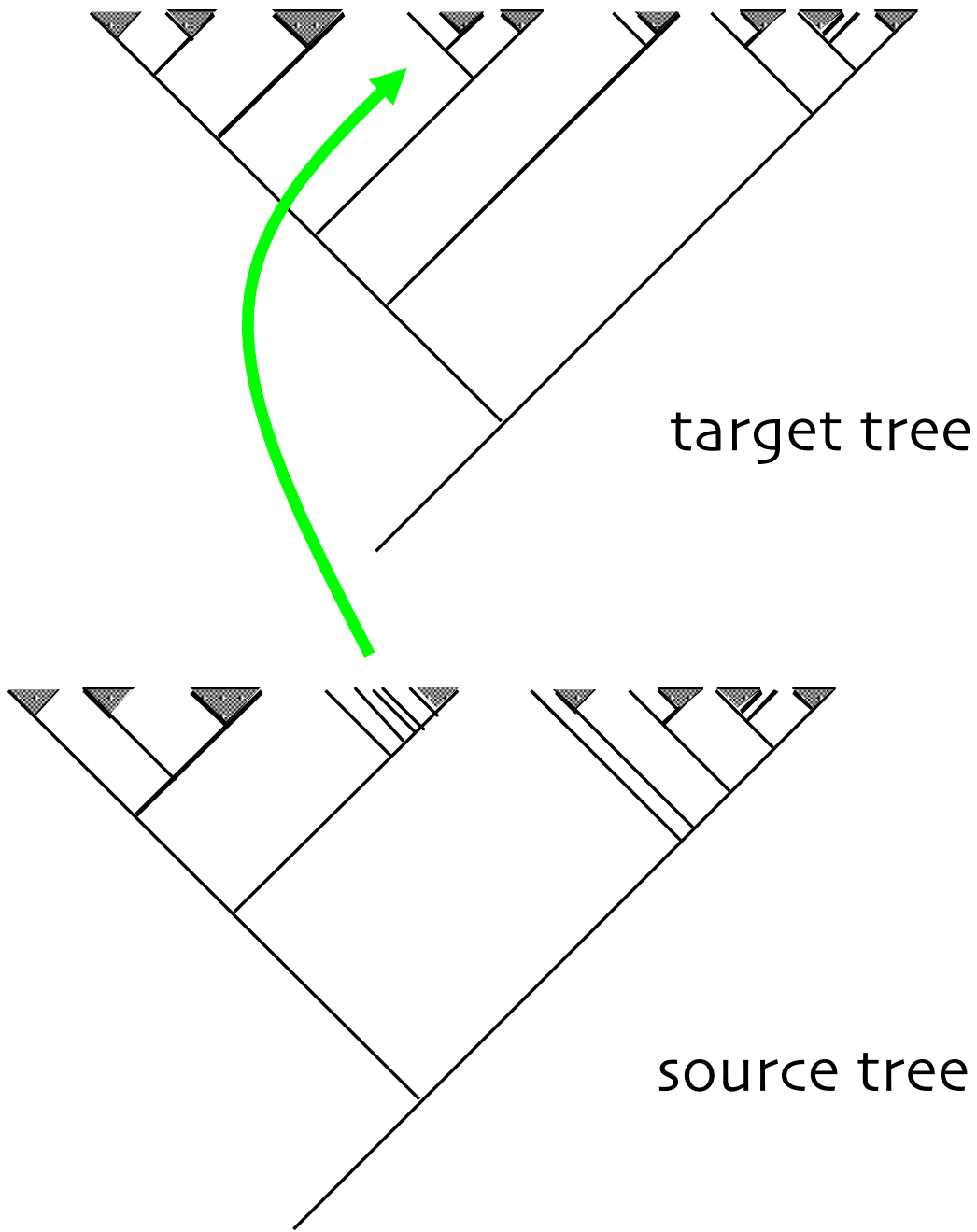
Mixed sectorial search (MSS)

1. initiated by using RAS+SPR
2. consensus of the shortest tree + shortest tree of previous search
3. consensus used as constraint for TBR algorithm
2x faster than unconstrained search
4. continued by using RSS

Goloboff, P. (& Moilanen, A.) Tree fusing

1. 2 starting trees chosen
2. trees compared one sector at a time
3. all sectors that reduce tree length transferred from source to target tree
4. a new source tree chosen

initially trees resulting from numerous searches needed
efficiency of the method based on the fact that at least
one part of the tree is in optimal configuration



target tree

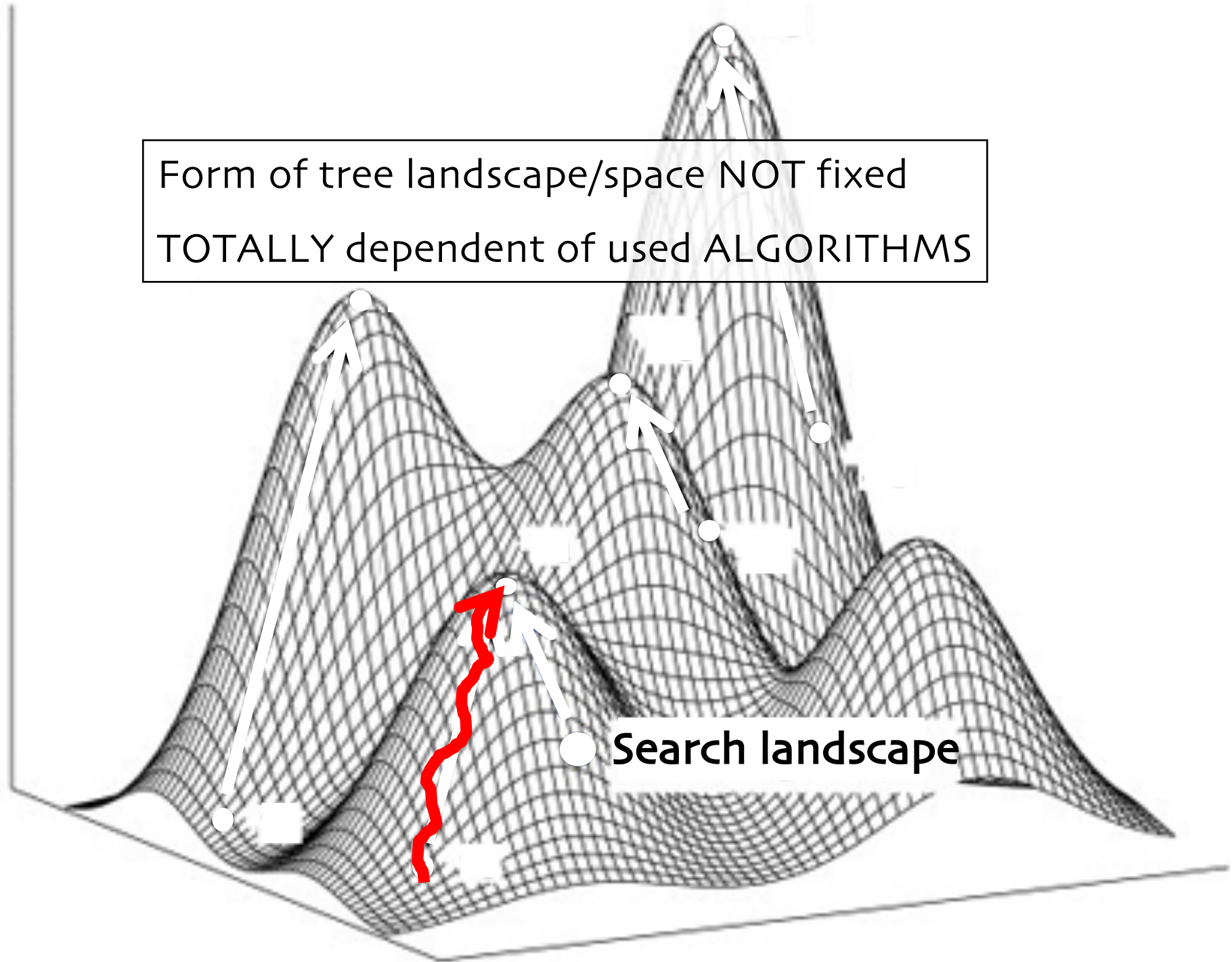
source tree

Goloboff, P. Tree drifting

longer than optimal trees accepted with predefined probability


widely known as “simulated annealing” used for analyses of difficult optimization problems

Form of tree landscape/space NOT fixed
TOTALLY dependent of used ALGORITHMS



Search landscape

Goloboff, P. A., Farris, J.S. & Nixon, K. TNT



use of novel strategies has led to PRONOUNCED decrease of time required for analyses and for more comprehensive analyses

zilla matrix (500 angiosperm *rbcl* sequences)

PAUP-analysis on Sun work stations >11 months,
shortest tree 16 220

TNT 200 MHz PC 4 min., shortest tree 16 218!!!

Phylogeny of the family Cladoniaceae (Lecanoromycetes, Ascomycota) based on sequences of multiple loci

Soili Stenroos^{a*†}, Raquel Pino-Boda^{b*†}, Jaakko Hyvönen^a, H. T. Lumbsch^c and Teuvo Ahti^a

^aFinnish Museum of Natural History, Botany Unit, University of Helsinki, PO Box 47, FI-00014, Helsinki, Finland; ^bReal Jardín Botánico de Madrid, CSIC, Plaza de Murillo 2, 28014, Madrid, Spain; ^cScience & Education, The Field Museum, 1400 S. Lake Shore Drive, Chicago, IL, 60605, USA

Accepted 15 October 2018

Abstract

Cladoniaceae is a family of lichenized fungi that belongs to the Lecanorales, Ascomycota. The family is distributed widely, although several genera are restricted to the Southern Hemisphere. The circumscriptions of the genera and species in the family have traditionally been based on thallus morphology, the type of vegetative propagules and the secondary metabolites. However, numerous species are highly variable phenotypically, making their delimitation problematic. In the present study a new phylogeny of Cladoniaceae is constructed using five loci (ITS rDNA, IGS rDNA, *RPB2*, *RPB1*, *EF-1a*) from a worldwide sample of 643 specimens representing 304 species. Cladoniaceae was resolved as a monophyletic group. The circumscription of the genera and the relationships among them are discussed. *Pycnothelia*, *Carassea* and *Metus* are closely related, forming a sister clade to the larger genus *Cladonia*. *Cladia* in its recent wide sense turned out to be paraphyletic, including species that have been recognized in *Thysanothecium* and *Notocladonia*. *Cladonia* was resolved as monophyletic, with *C. wainioi* as the earliest diverging lineage. Eleven major clades were resolved in *Cladonia*. No synapomorphies were found for most of them. We propose the new genera *Pulchrocladia* and *Rexia*, as segregates of *Cladia*, five new combinations, and the resurrection of the genus *Heterodea*.

12363	WILEY	Dispatch: 9.11.18	CE: Raja S
		No. of pages: 34	PE: Megala R.
Manuscript No.			

Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups

Pablo A. Goloboff^{a,*}, Santiago A. Catalano^b, J. Marcos Mirande^b, Claudia A. Szumik^a,
J. Salvador Arias^a, Mari Källersjö^c and James S. Farris^d

^aINSUE (Instituto Superior de Entomología), CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), Instituto Miguel Lillo, Miguel Lillo 205, 4000 S.M.Tucumán, Argentina; ^bCONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), Fundación Miguel Lillo, Miguel Lillo 251, 4000 S.M.Tucumán, Argentina; ^cGöteborgs Botaniska Trädgård (Gothenburgh Botanical Garden), Carl Skottbergs Gata 22A, SE-413 19 Göteborg, Sweden; ^dMolekylärsystematiska laboratoriet, Naturhistoriska riksmuseet, Box 50007, 104-05 Stockholm, Sweden

Accepted 21 February 2009

Abstract

Obtaining a well supported schema of phylogenetic relationships among the major groups of living organisms requires considering as much taxonomic diversity as possible, but the computational cost of calculating large phylogenies has so far been a major obstacle. We show here that the parsimony algorithms implemented in TNT can successfully process the largest phylogenetic data set ever analysed, consisting of molecular sequences and morphology for 73 060 eukaryotic taxa. The trees resulting from molecules alone display a high degree of congruence with the major taxonomic groups, with a small proportion of misplaced species; the combined data set retrieves these groups with even higher congruence. This shows that tree-calculation algorithms effectively

Concluding remarks

solutions to find optimal trees varies according to the data analyzed

e.g. extensive homoplasy vs. **randomly sparse** matrices

increasingly common with large genetic materials

Wagner algorithm used initially to find starting trees
modified selected/informative addition

Goloboff, P.A. 2014. Hide and vanish: data sets where the most parsimonious tree is known but hard to find, and their implications for tree search methods. *Molecular Phylogenetics & Evolution* 79: 118-131.

PROGRAMS AVAILABLE

Mesquite

nona

+ winclada

www.cladistics.org

www.lillo.org.ar/phylogeny/

<http://evolution.genetics.washington.edu/phylip/software.html>

PARALLELIZATION of programs

- problems divided into smaller parts > distributed to SEVERAL CPUs to be solved SIMULTANEOUSLY

Evaluating results

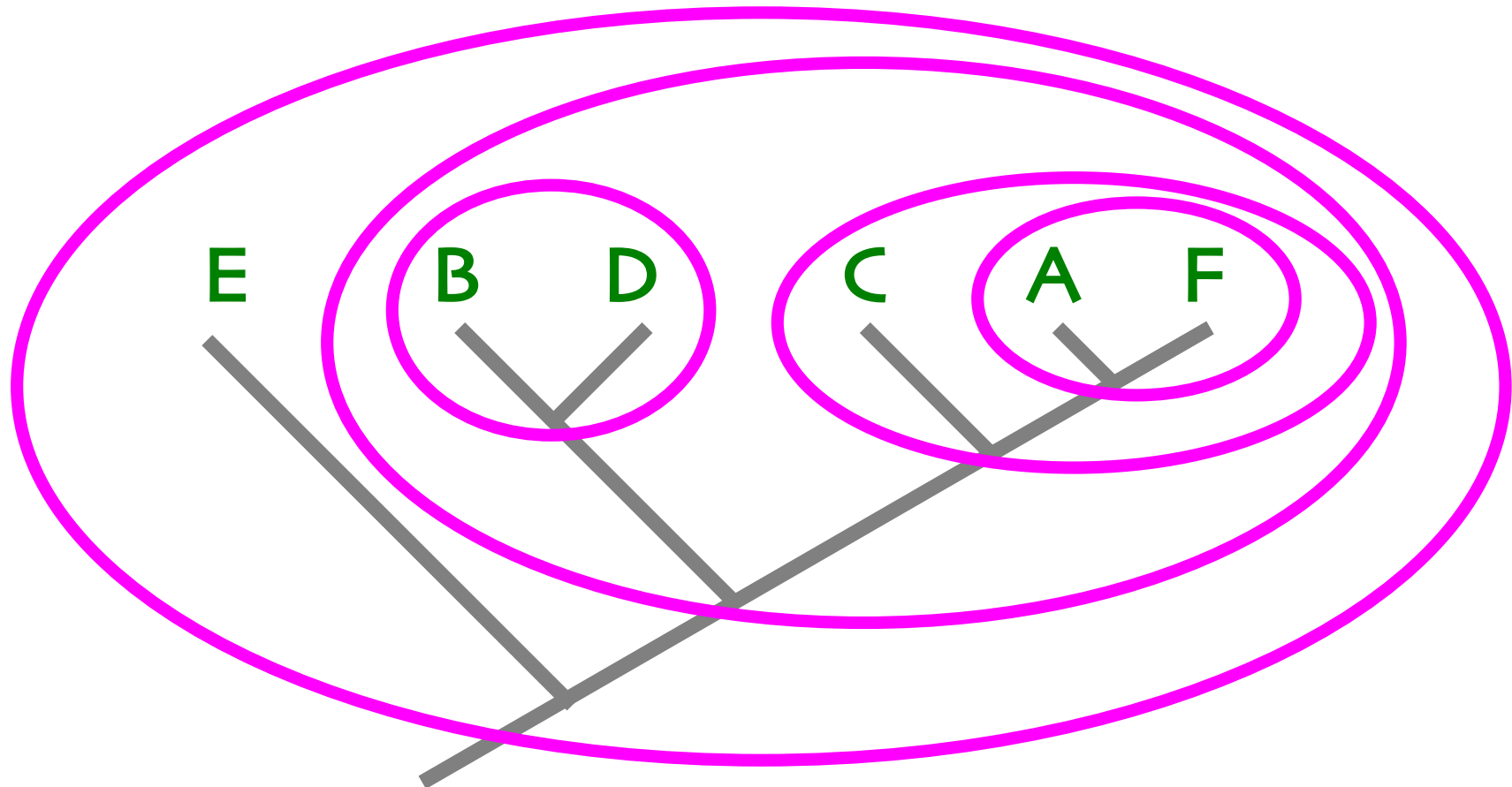


can we trust the results obtained?

are part of the results simply accidental?

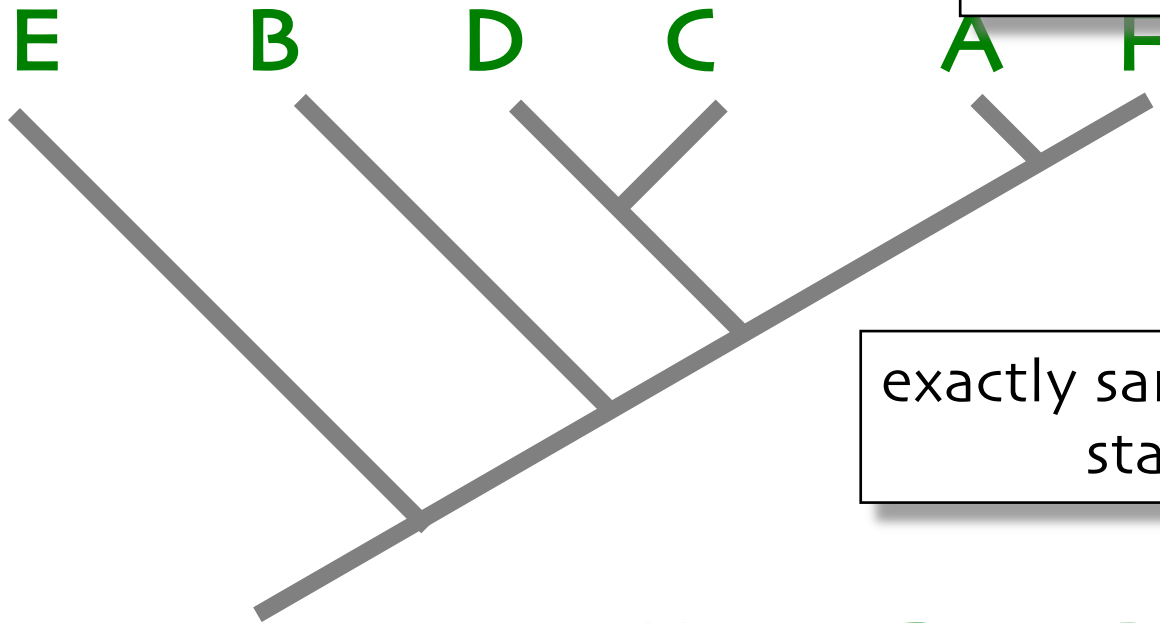
which PARTS of tree are most reliable?

MONOPHYLY, paraphyly, polyphyly

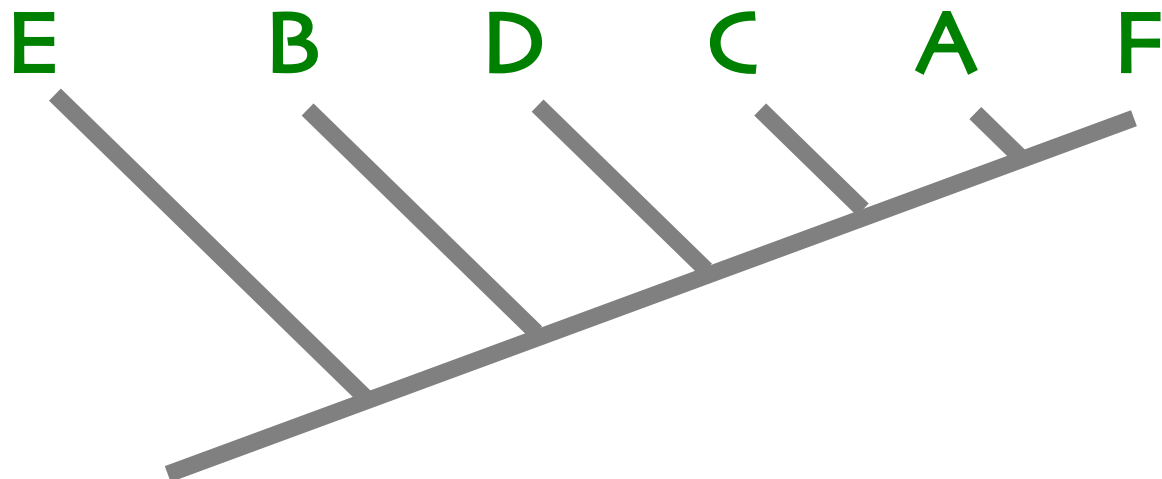


Numerous equally parsimonious trees

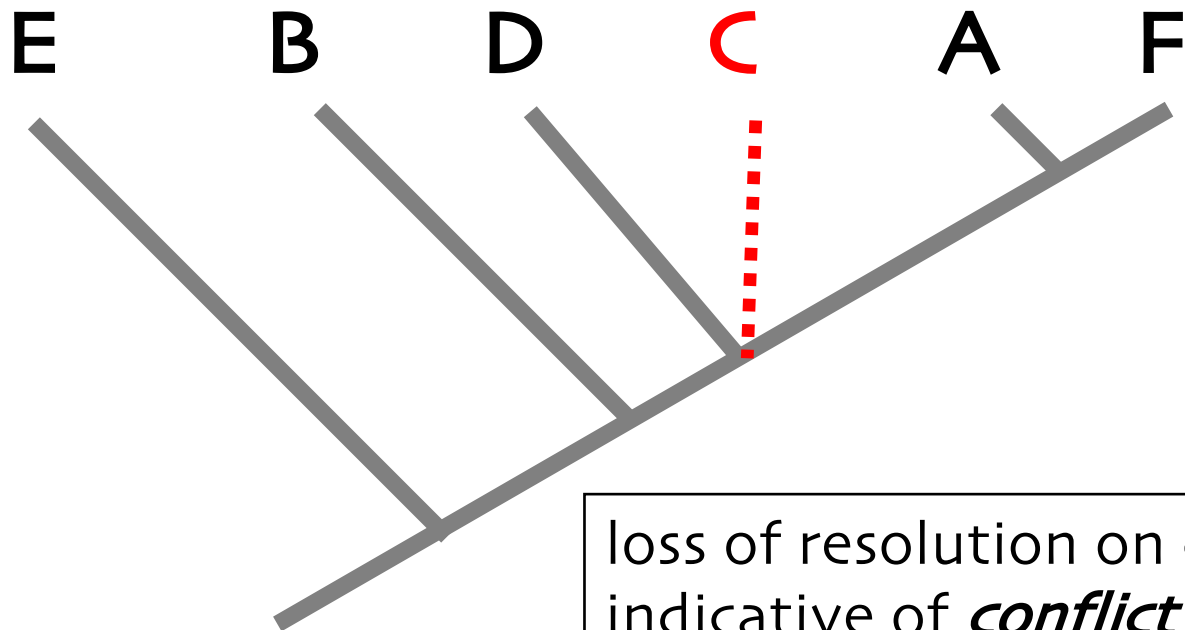
EQUALLY simple explanations
for data



exactly same number of ch.
state changes



Consensus tree



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

Evaluating results

support values

3 commonly used methods:

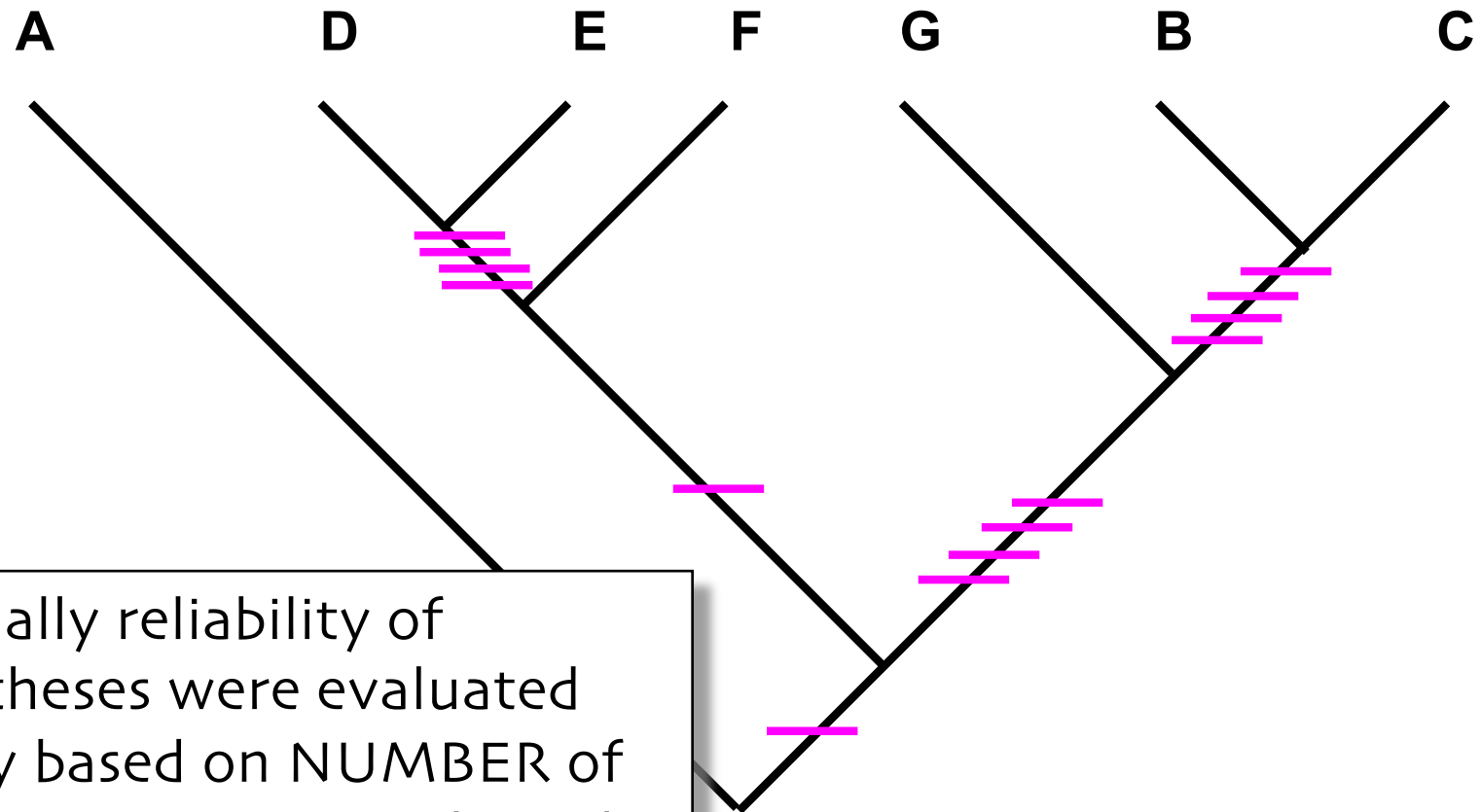
Bremer support value

branch support

Bootstrap

Parsimony jackknifing

Bremer support



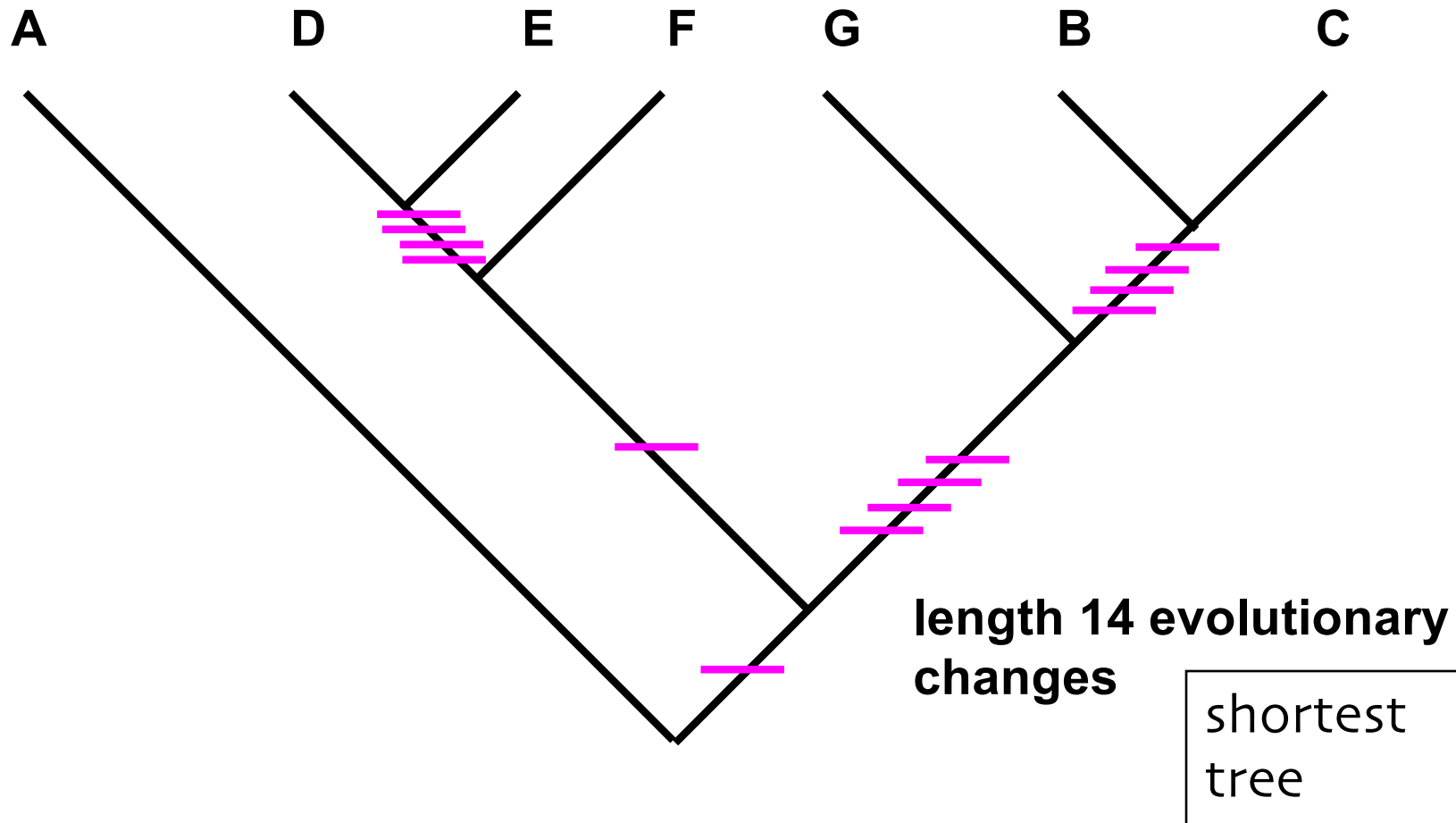
originally reliability of hypotheses were evaluated purely based on NUMBER of characters supporting branch

Bremer support

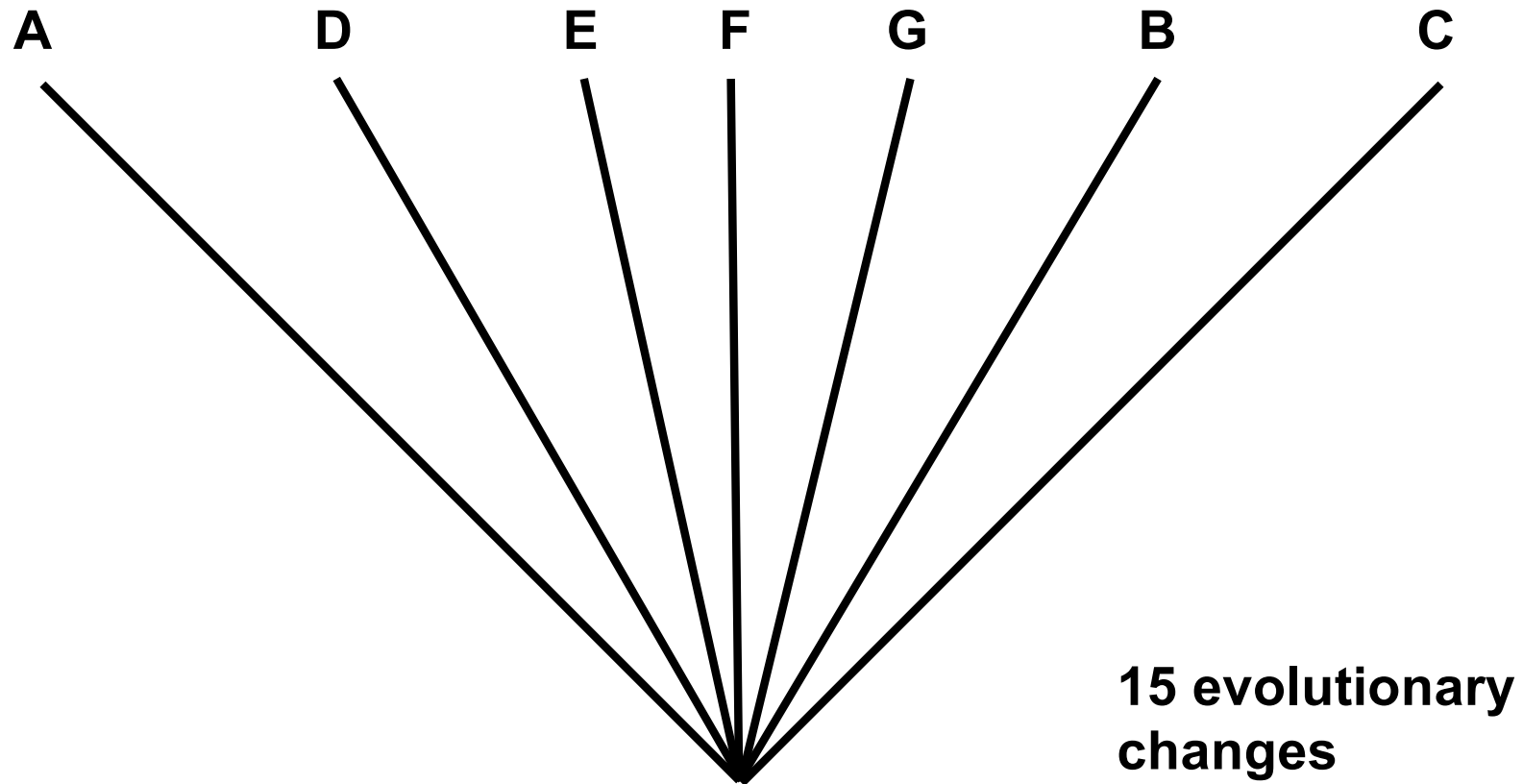
	0	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	0	0	0	0	0
B	1	0	1	1	1	1	1	1	1	1
C	1	0	1	1	1	1	1	1	1	1
D	1	1	0	0	0	0	1	1	1	1
E	1	1	0	0	0	0	1	1	1	1
F	1	1	0	0	0	0	0	0	0	0
G	1	0	1	1	1	1	0	0	0	0

7 x 10

Bremer support



Bremer support



consensus of trees with length ≤ 15

TAXONOMIC CHARACTERS

characters used in phylogenetic analyses are assumed to be **INDEPENDENT** of other characters

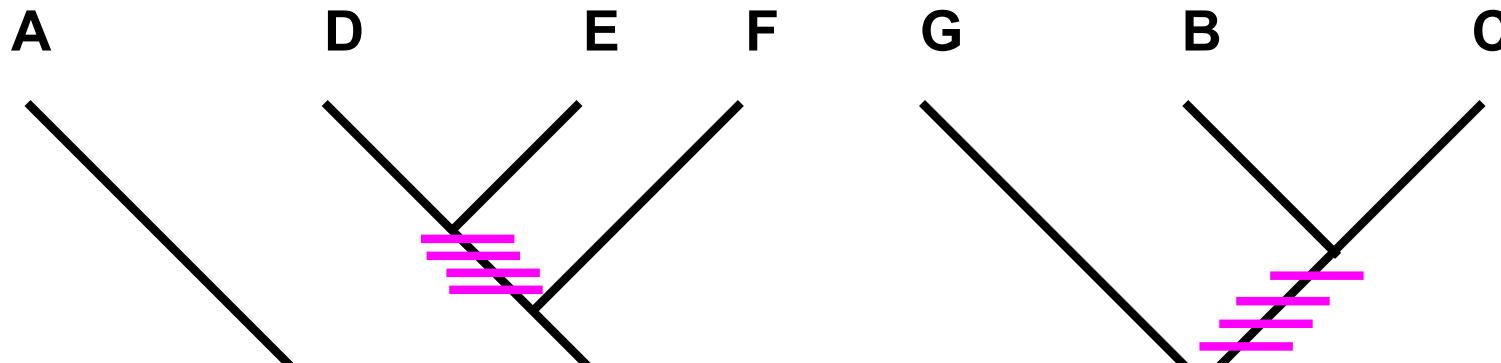
NO genetic correlation

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

but characters **DO INTERACT** within the matrix

Bremer s

despite of the LARGE differences in synapomorphies of monophyletic groups ALL of these lost on consensus based on trees ≤ 15



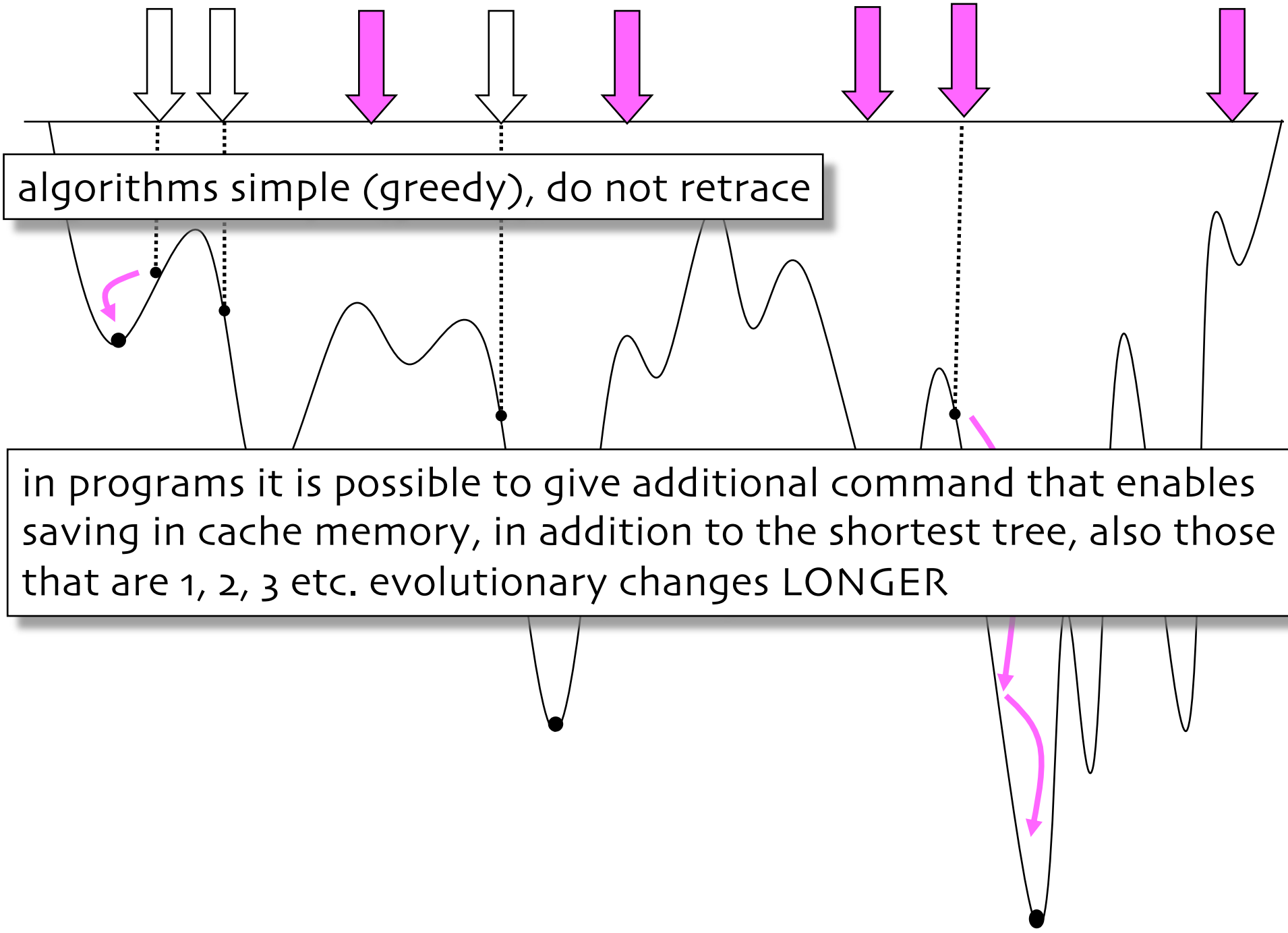
Goodman, M. & al. 1982. New perspectives in the molecular biological analysis of mammalian phylogeny. *Acta Zoologica Fennica* 169: 19-35.

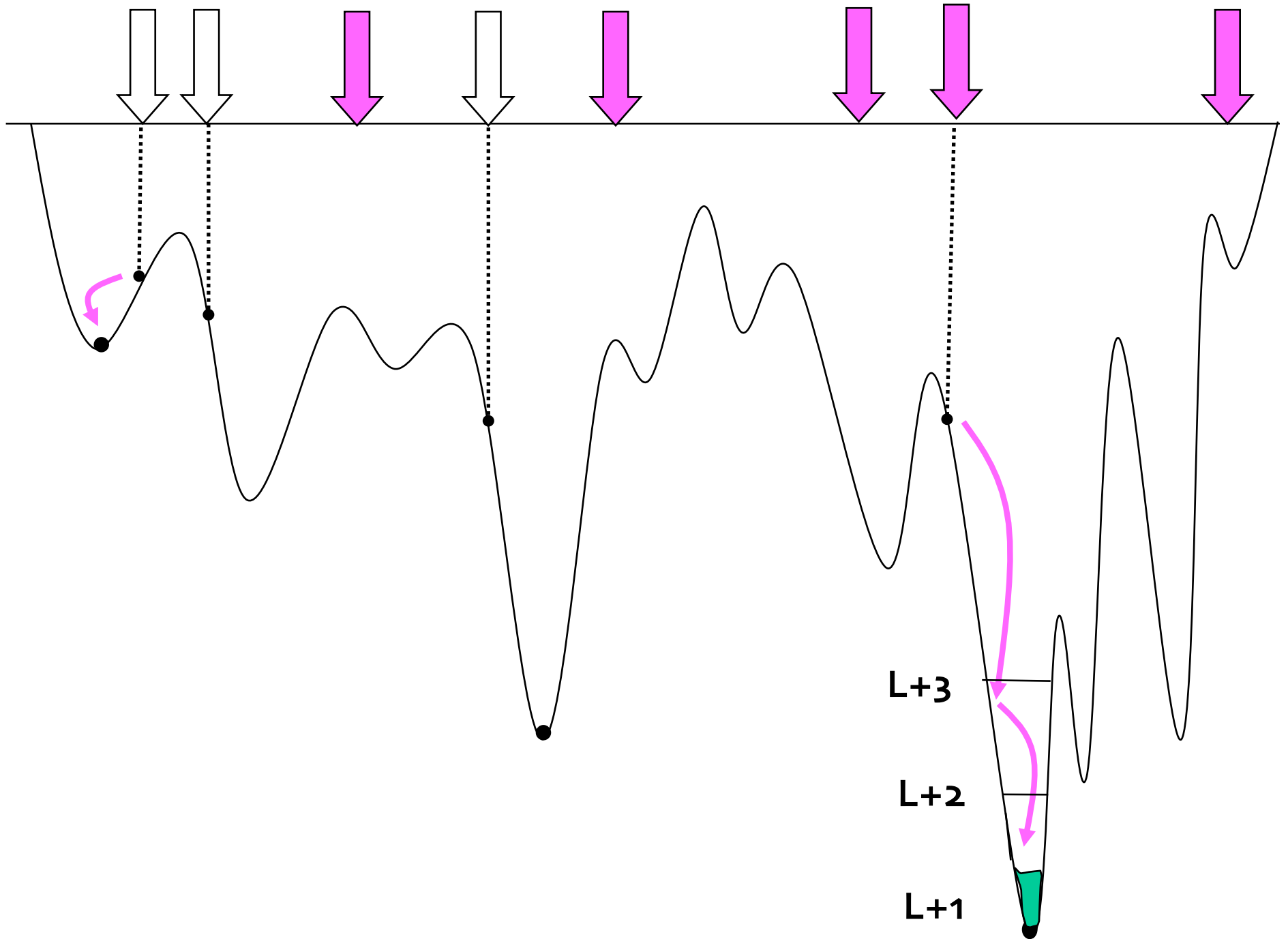
Goodman-Bremer support
decay index

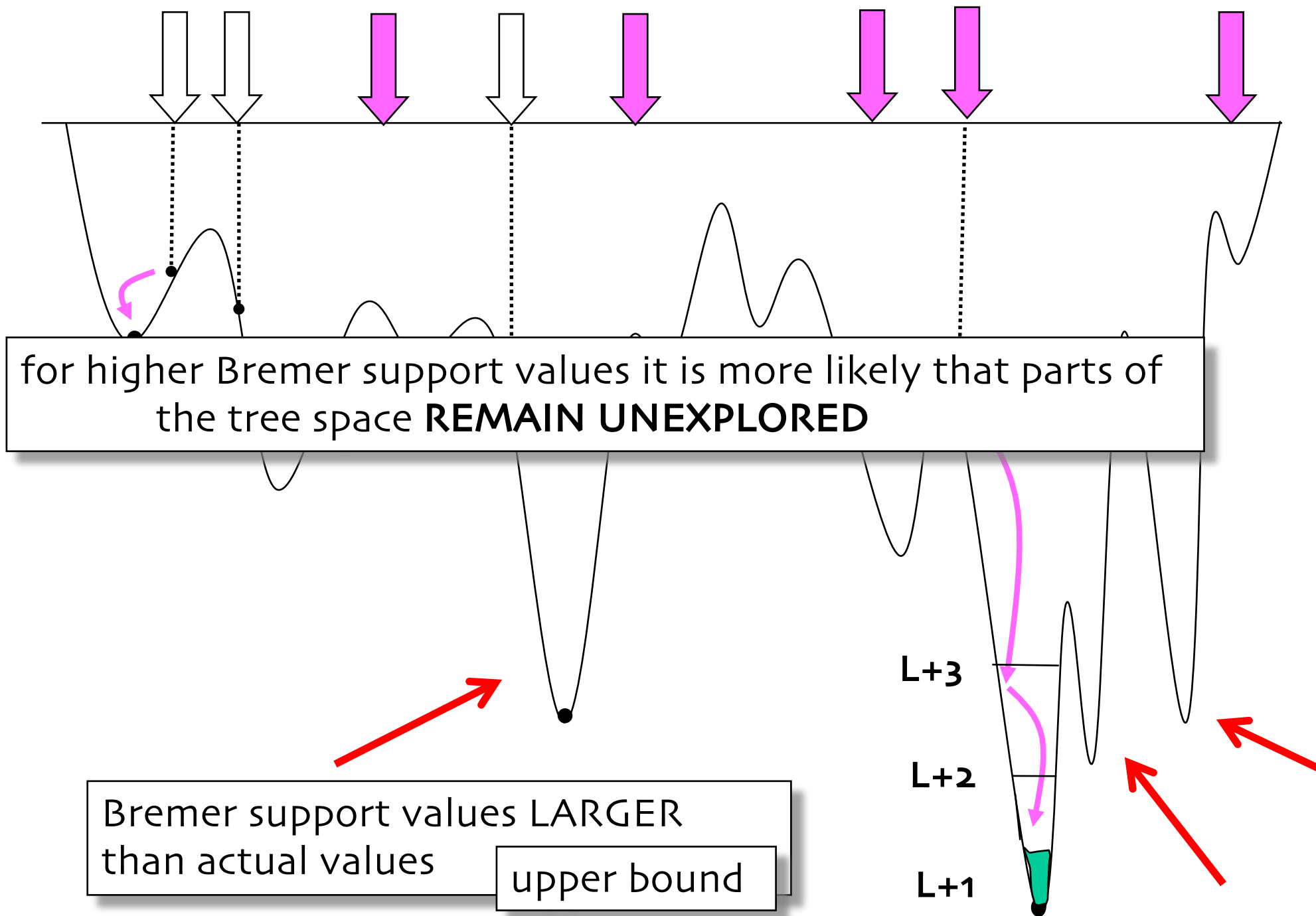
Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42: 795-803.

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

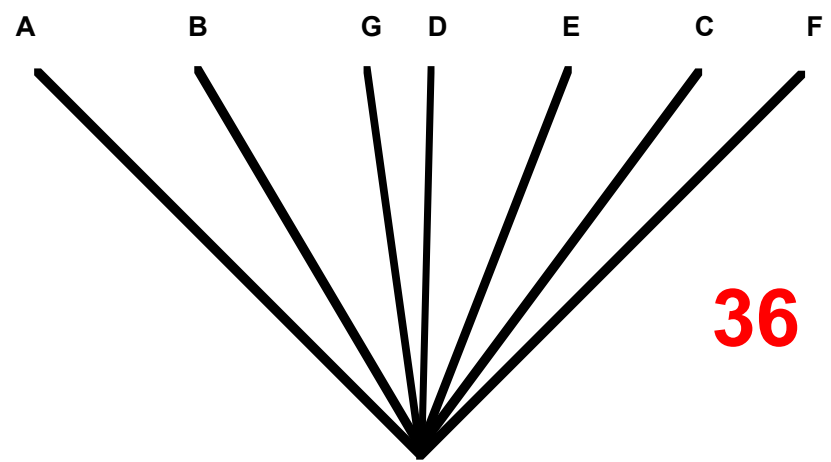
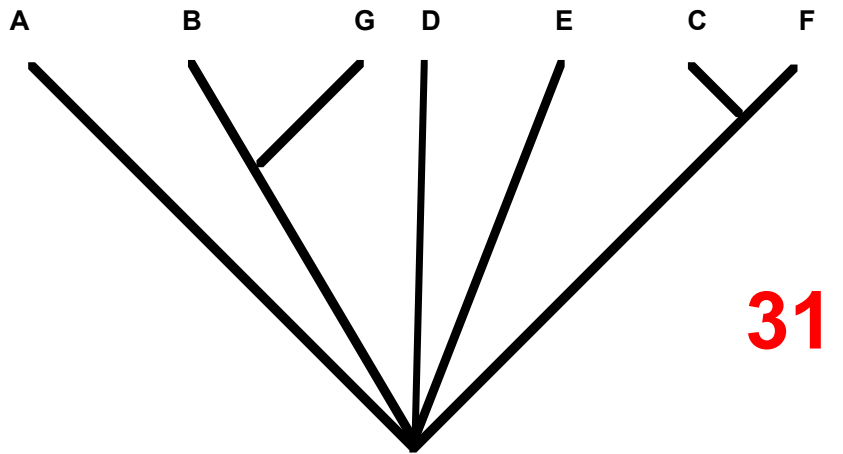
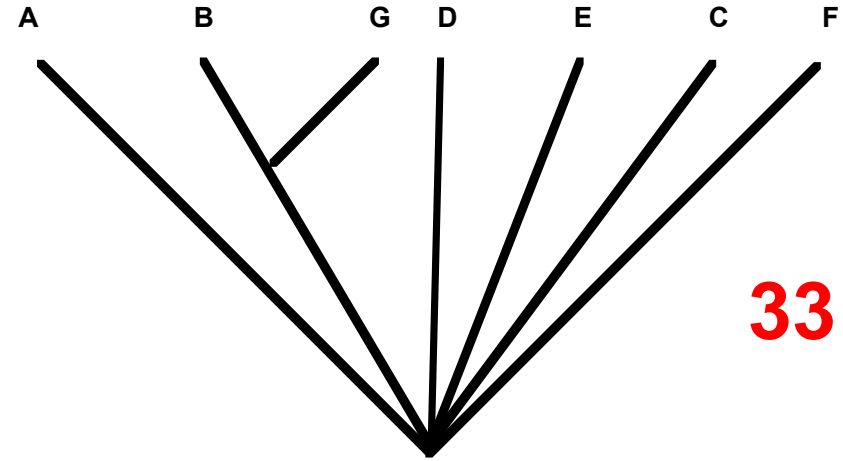
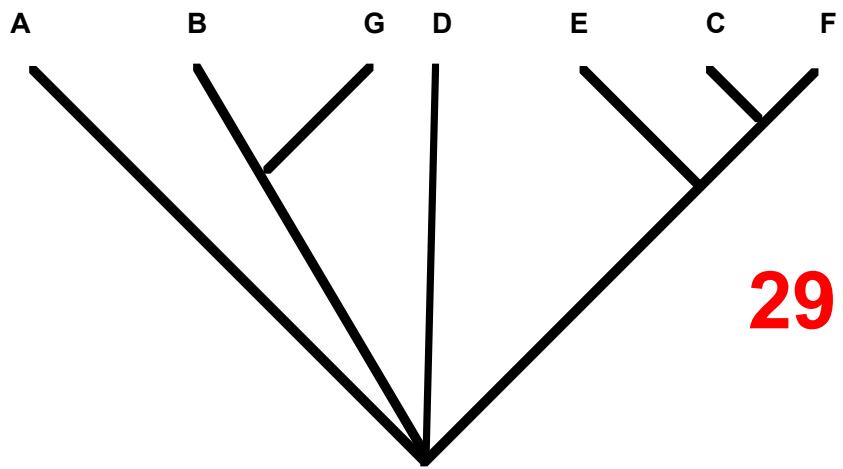
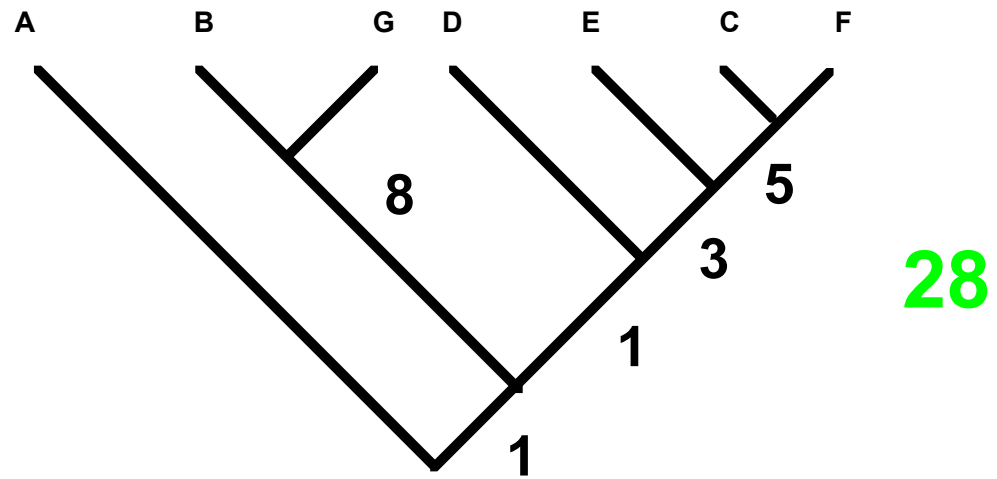




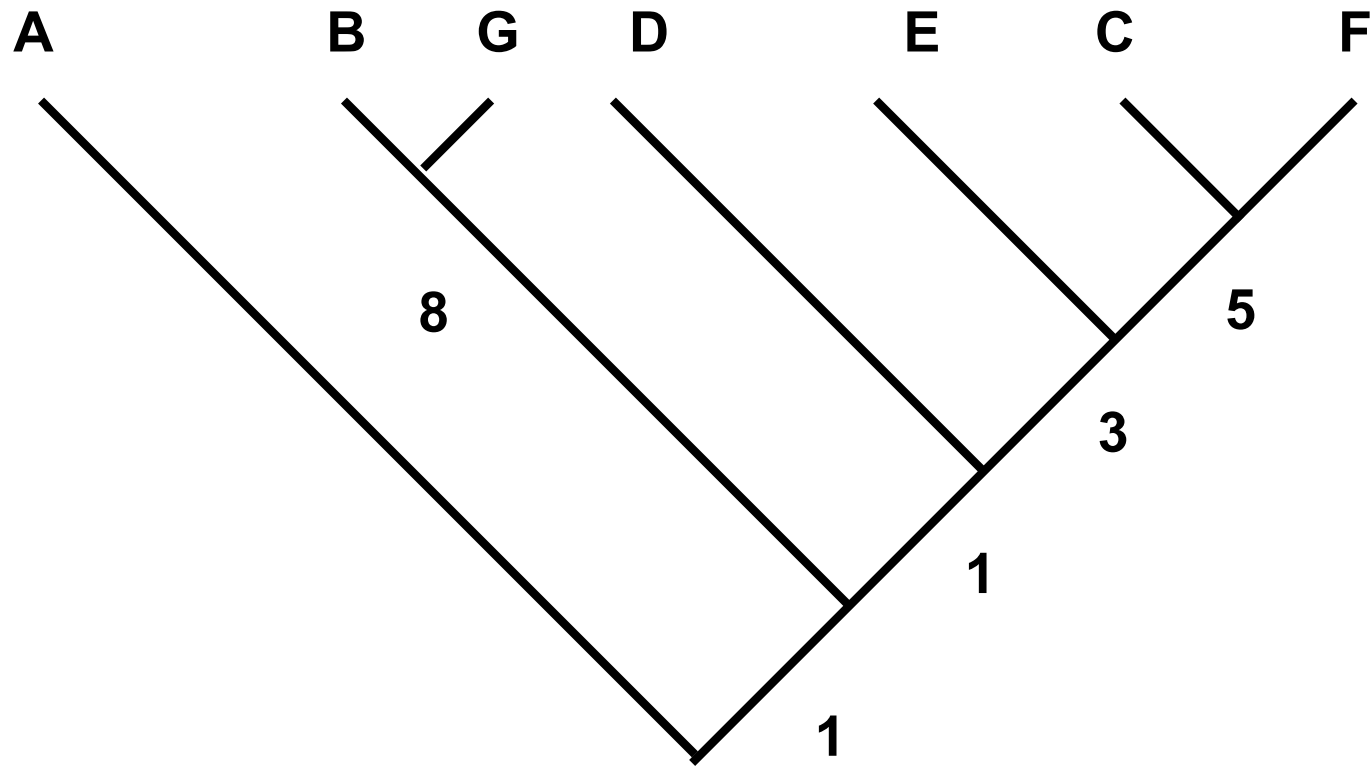


Bremer support

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)



Bremer support



18

Bremer support value for whole tree

Bremer support

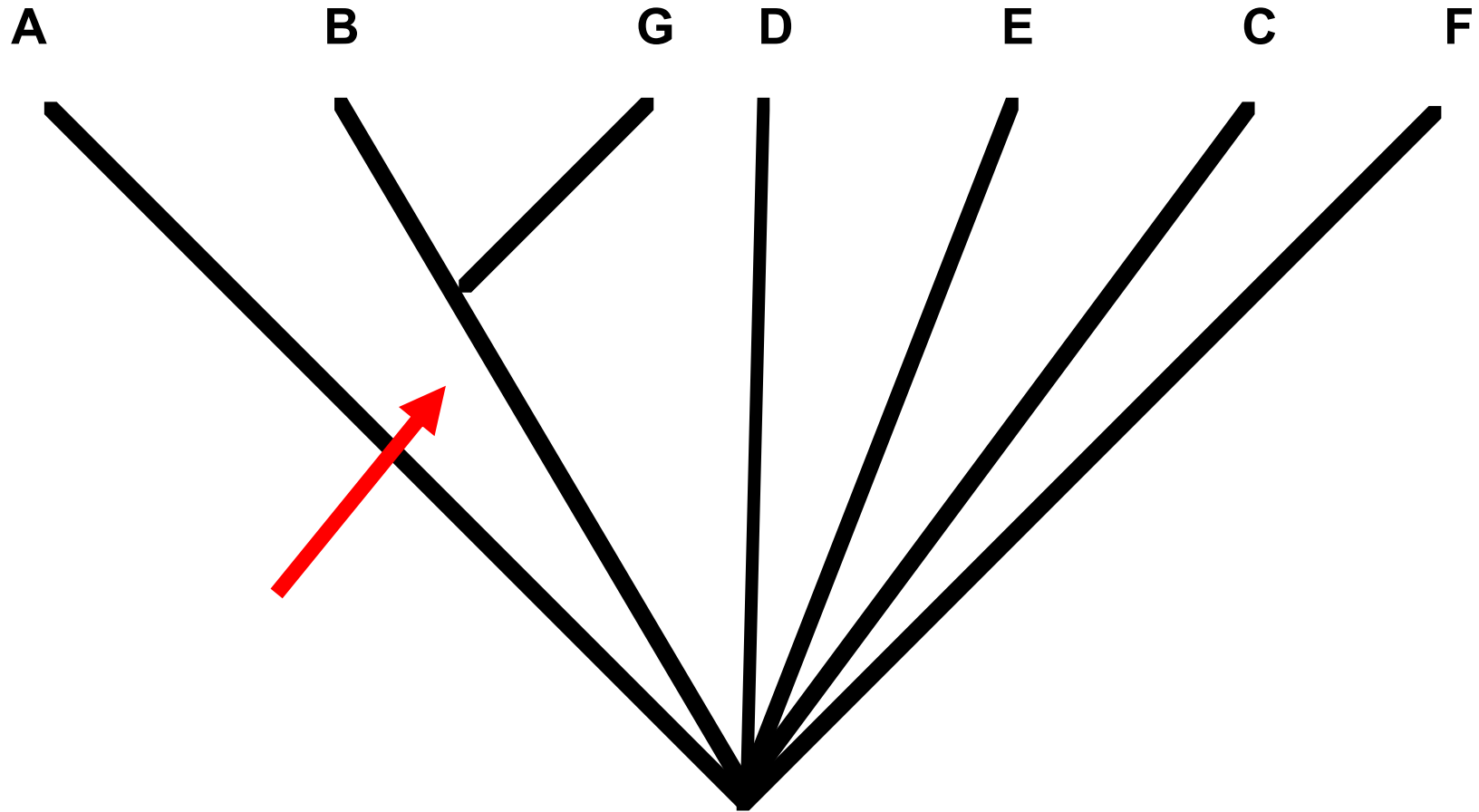
Bremer support value can be calculated also for *individual parts of tree* using constrained search

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)

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

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

F = synapomorphies of the group inspected
C = synapomorphies of groups in CONFLICT with the group inspected

$$0 < \text{RFD} < 1$$

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 res



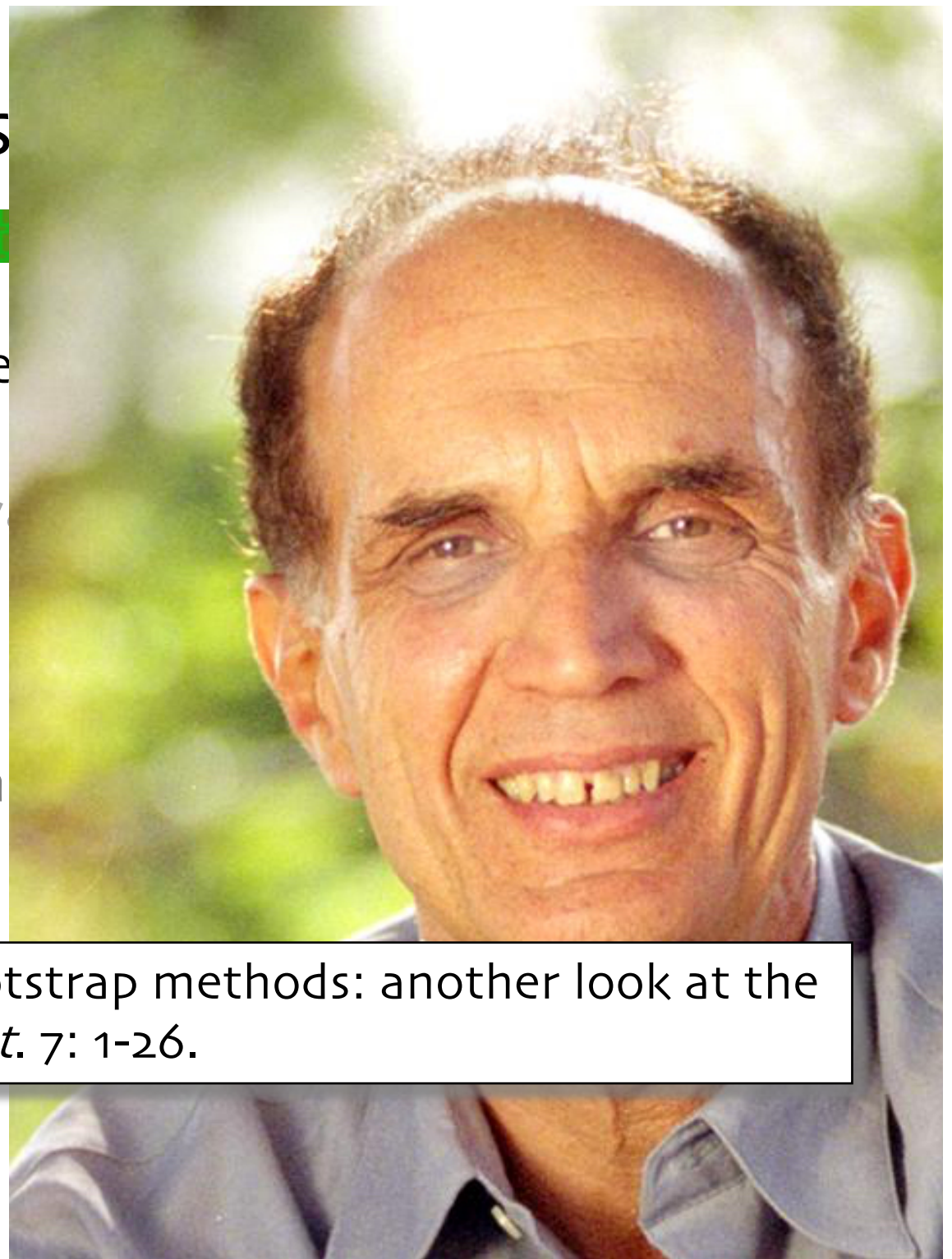
3 commonly used me

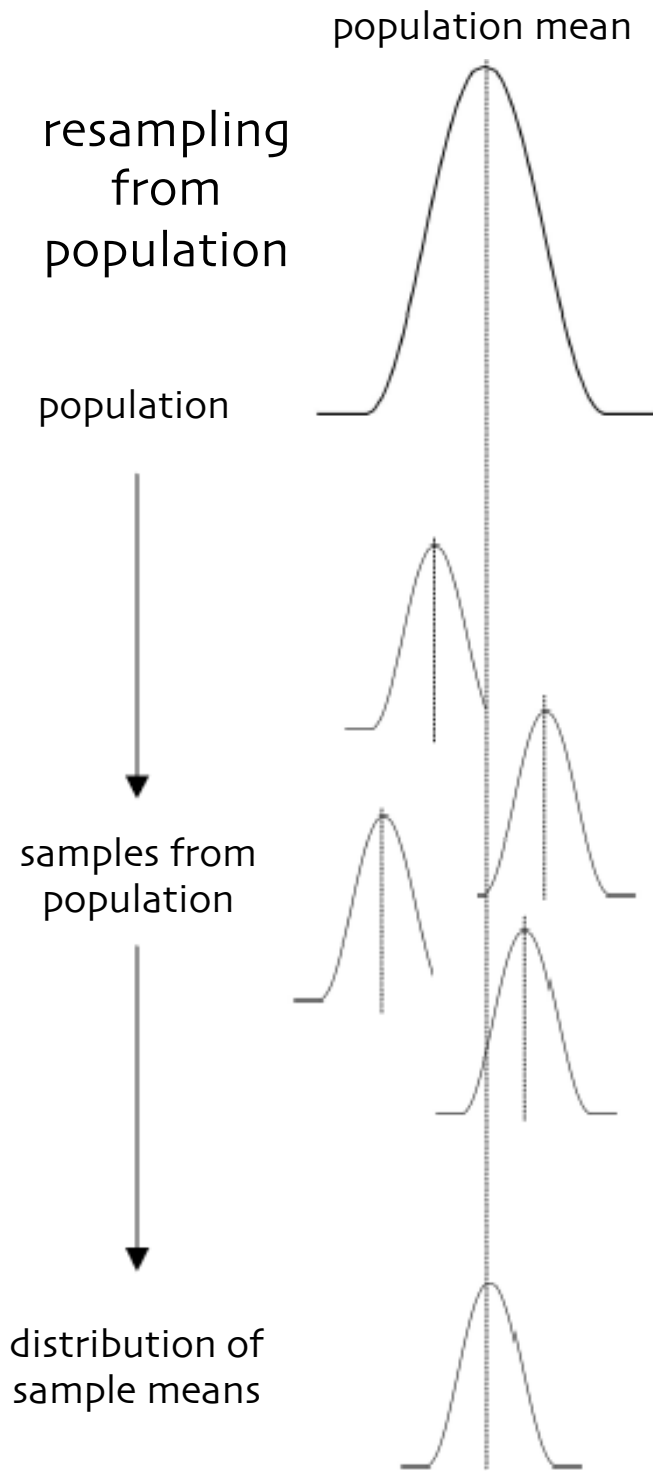
Bremer support v

Bootstrap

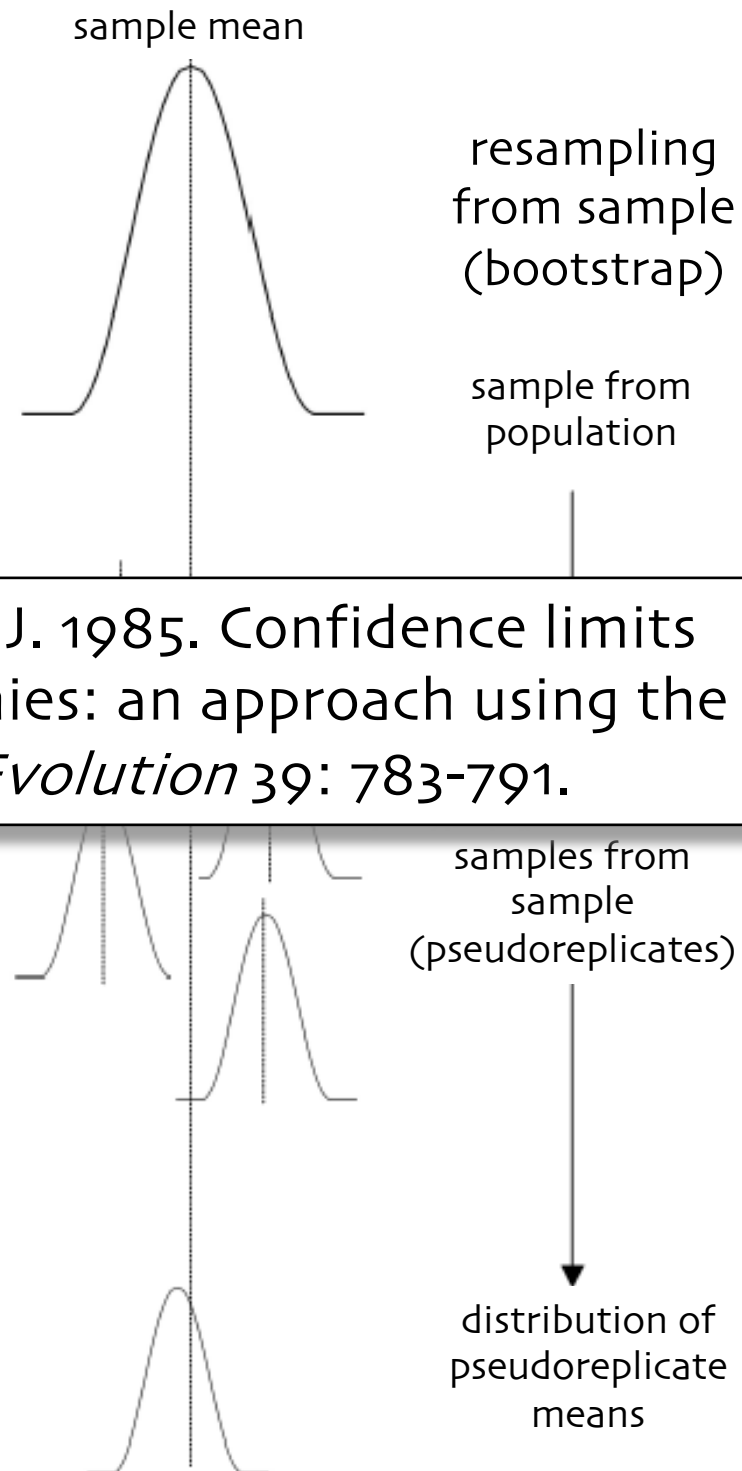
Parsimony jackkn

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

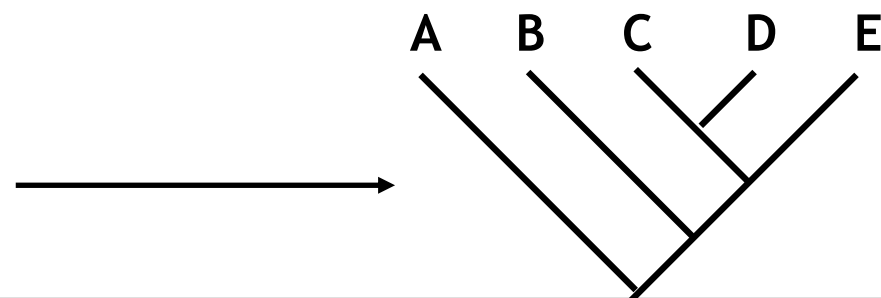




Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791.



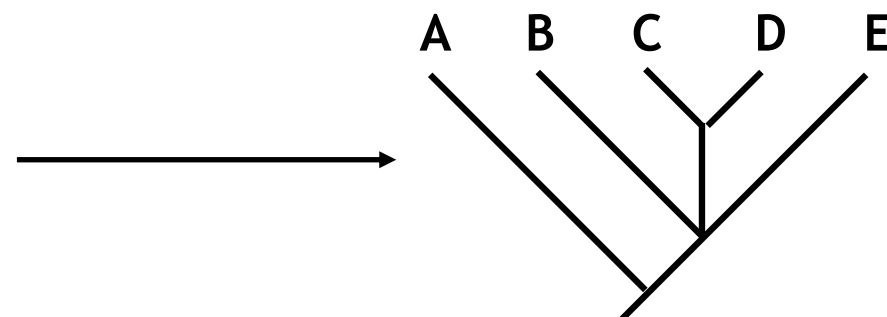
	characters									
taxa	0	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	1	0	0	1	0
B	1	1	0	0	0	0	0	0	0	0
C	1	1	1	1	1	1	1	2	0	1
D	1	1	1	1	0	1	1	2	0	1
E	1	1	0	0	1	0	1	1	1	0



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!)

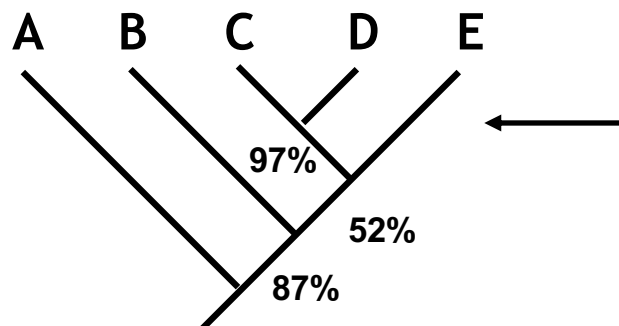
	characters									
taxa	7	8	3	7	4	5	3	0	1	9
A	0	0	0	0	0	1	0	0	1	0
B	0	1	0	0	0	0	0	0	0	0
C	2	1	1	1	1	1	1	2	0	1
D	2	1	1	1	0	1	1	2	0	1
E	1	1	0	0	1	0	1	1	1	0



repeated several times (100- 10 000 x)

results combined to a majority rule compromise tree

disadvantage: autapomorphies, invariable characters affect the values



BOOTSTRAP

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:

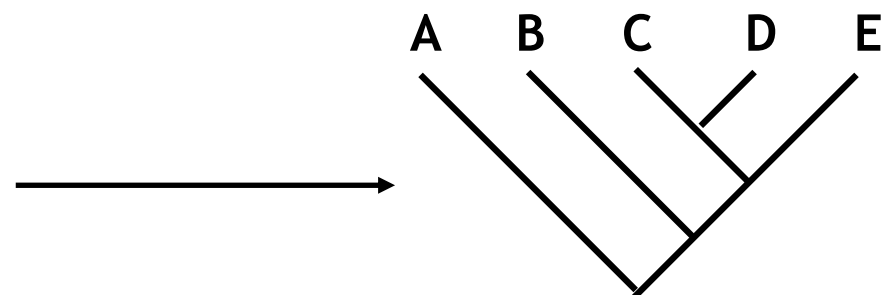
Bremer (branch) support

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.

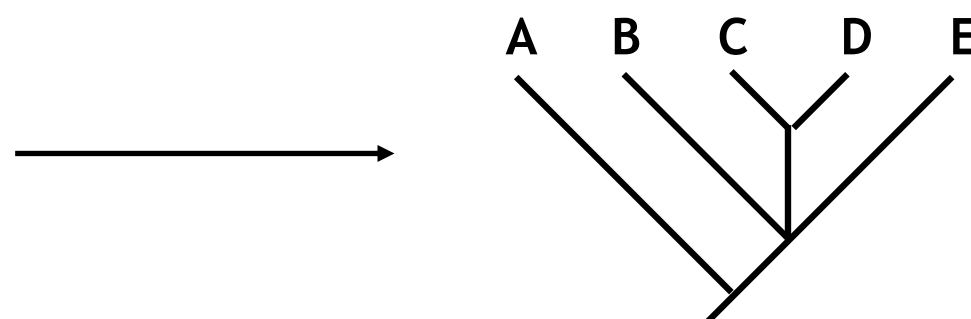
	characters									
taxa	0	1	2	3	4	5	6	7	8	9
A	0	0	0	0	0	1	0	0	1	0
B	1	1	0	0	0	0	0	0	0	0
C	1	1	1	1	1	1	1	2	0	1
D	1	1	1	1	0	1	1	2	0	1
E	1	1	0	0	1	0	1	1	1	0



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

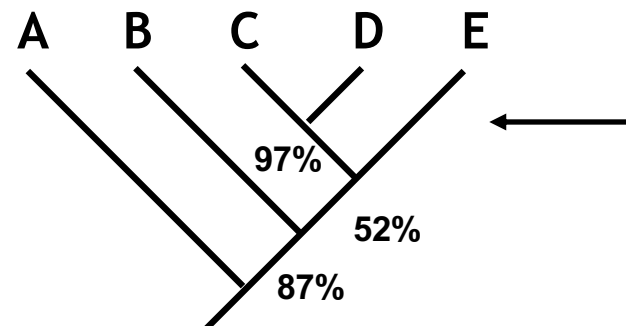
e.g. sampling is made so that for a single character probability of remaining UNSAMPLED is $1/e$ (~37%)

	characters						
taxa	0	1	3	4	7	8	9
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



repeated several times (100- 10 000 x)

results combined to a majority rule compromise tree



**PARSIMONY
JACKKNIFING**

Evaluating results



can we trust the results obtained?

are part of the results simply **accidental**?

which PARTS of tree are most reliable?

Evaluating results

several indices have been proposed for finding out whether available data deviates from that obtained by chance alone

e.g. PTP, cladogram length skewness

implicitly & superficially appealing approaches

unfortunately ONLY able to tell that the data is not accidental, i.e. also data WITHOUT any phylogenetic signal will get significant values

i.e. data with such character incongruence & internal conflict that no phylogenetic data seem to be present

Carpenter, J.M. 1992. Random cladistics. *Cladistics* 8: 147-153.

Evaluating results

one goal is to estimate how easily the obtained results, i.e. tree (or its parts) will change if we add new characters into our matrix

all indices given above are INDIRECT ways to estimate this

we do NOT know this BEFORE a new analysis is made

part of the added new characters are congruent, part in conflict with presented results

different support values give in many cases comparable results, same groups revealed

Evaluating results



same indices used for analyses using DIFFERENT
optimality criteria

how they behave with these DIFFER

Simmons, M.P. & Goloboff, P.A. 2014. Dubious resolution and support from published sparse supermatrices: the importance of thorough tree searches. *Molecular Phylogenetics & Evolution* 78: 334-348.

Evaluating results



GO AND GET MORE DATA

only NEW characters will be able to
REALLY evaluate (test)
results obtained

Grant, T. & Kluge, A.G. 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. *Cladistics* 19: 379-418.

SIMULTANEOUS ANALYSIS

ALL characters of ALL stages of life-cycle should be combined into a same matrix for analysis

WHY?

by including into our analysis characters simultaneously we “test” them against each other

the more characters we have in our analysis, the more severe is our “test”

more chances for characters to be in conflict

Evaluating results

	1	2	3	4	5	6	7	8	9	10	11	12
OG	0	0	0	0	0	0	0	0	0	0	1	0
A	1	0	0	0	1	1	0	0	0	1	1	0
B	1	0	0	1	0	0	0	0	0	1	2	1
C	0	0	0	0	0	0	0	0	1	1	1	2
D	0	1	1	0	0	0	1	1	0	1	0	1
E	0	1	1	1	0	0	0	1	0	1	0	1

new cladistic analysis
do results change?
if yes, which parts?

Evaluating results

our goal is to estimate PHYLOGENY, history of lineages

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 littered with unlikely events

applicable also to phylogeny

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

SUMMARY

use of best programs & efficient algorithms
necessary for analyses of LARGE matrices

PARALLELIZATION have enabled analyses of larger
and larger materials

to be continued...

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

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!