16.xi.

1. compromise trees (continued...)

2. simultaneous analysis

3. search algorithms

4. traditional search strategies

5. new search strategies
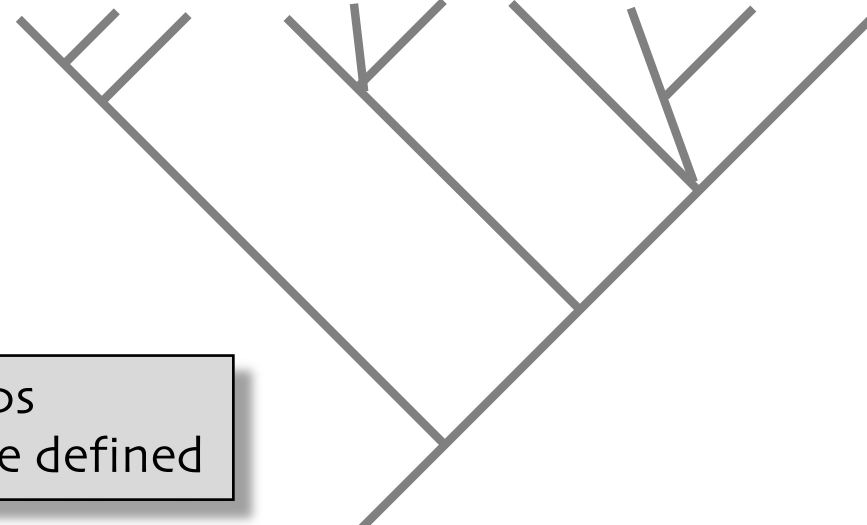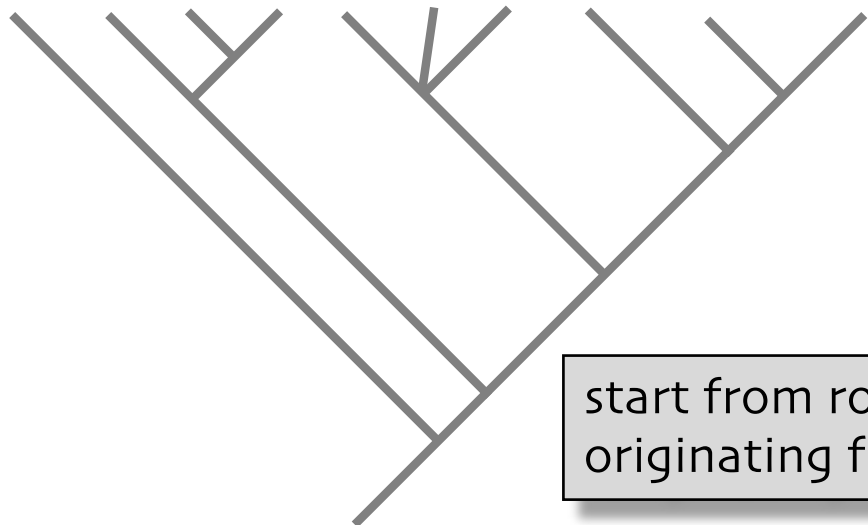
6. summary

COMPROMISE TREES

Adams, E. N. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology* 21: 390-397

Adams compromise   Adams II

J I H G F E D C B A   J I H G F E D C B A

start from root, groups
originating from there defined

(J) (I H G F E D C B A)   (J I H)(G F E D C B A)

search for non-empty intersections between groups, these marked at base
of tree if they include ≤ 2 terminals

(J)(I H)(G F E D C B A)

proceed to groups not yet on the compromise tree and treat them as above
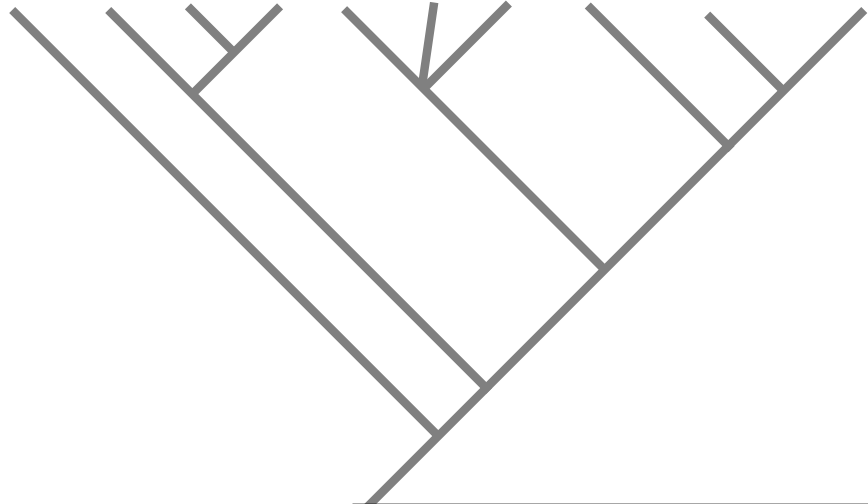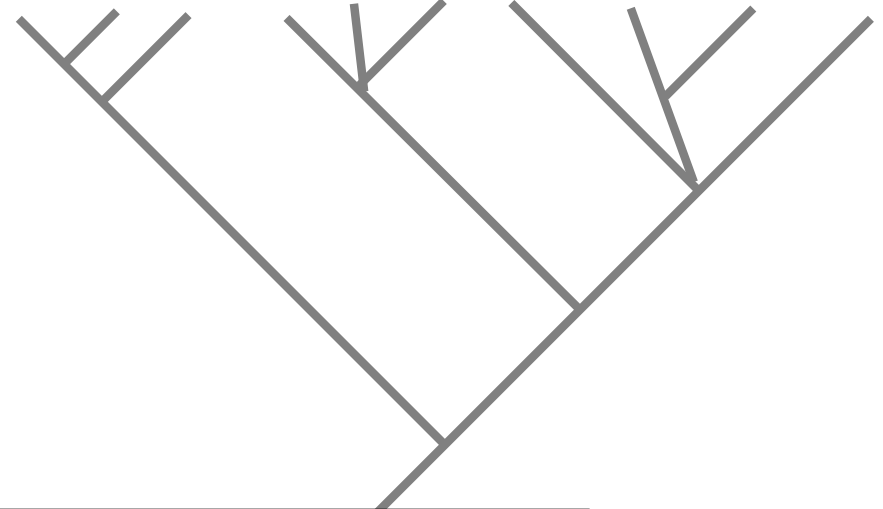
(G)(F E D C B A)   (G F E)(D C B A)

(G)(F E)(D C B A)

(D)(C B A)   (D C B)(A)

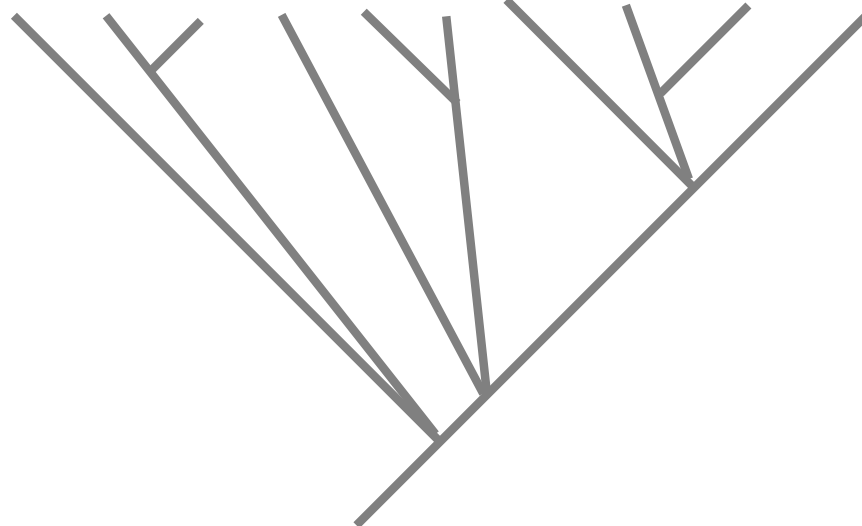J I H G F E D C B A

J I H G F E D C B A

(H I) monophyletic on compromise tree despite the fact that it is PARAPHYLETIC on both original trees

J I H G F E D C B A

E B D C A F   E A B D C F

Adams compromise

E A B D C F
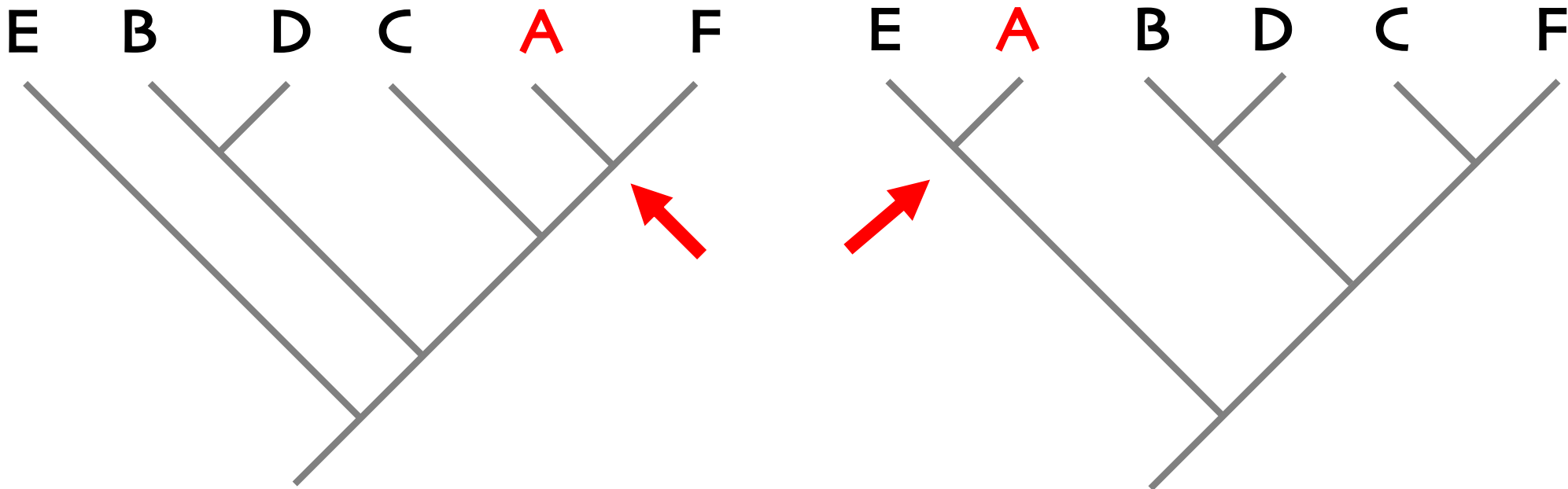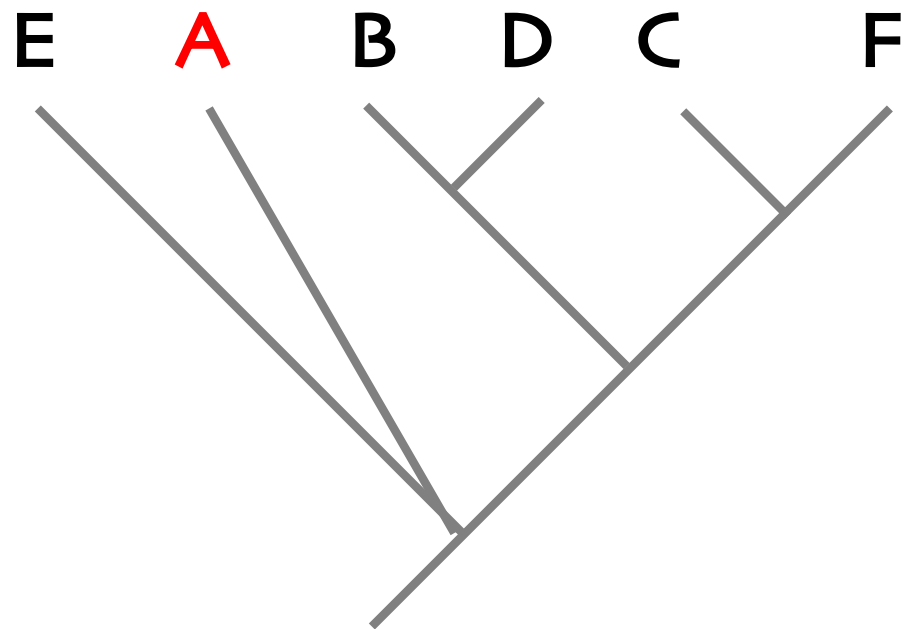
COMPROMISE TREES

Adams compromise

terminals causing conflict on compromise tree placed on positions that is common to all original trees

highlights **problematic** terminals

wild card taxa

might lead to presentation of groups ABSENT from original trees!

COMPROMISE TREES

majority rule

Adams

combinable component (semistric

Bremer, K. 1990. Combinable component consensus. *Cladistics* 6: 369-372.

COMPROMISE TREES

combinable component (semistrict)

all groups presented that are NOT in conflict with the optimal tree

this kind of compromise tree might include groups that are impossible to be present SIMULTANEOUSLY

1   O   O   O   1   1

A   B   C   D   E   F

1 > 0

1 > 0

combinable component (semistrict)

A B C D E F   A B C D E F

1 > 0

0 > 1
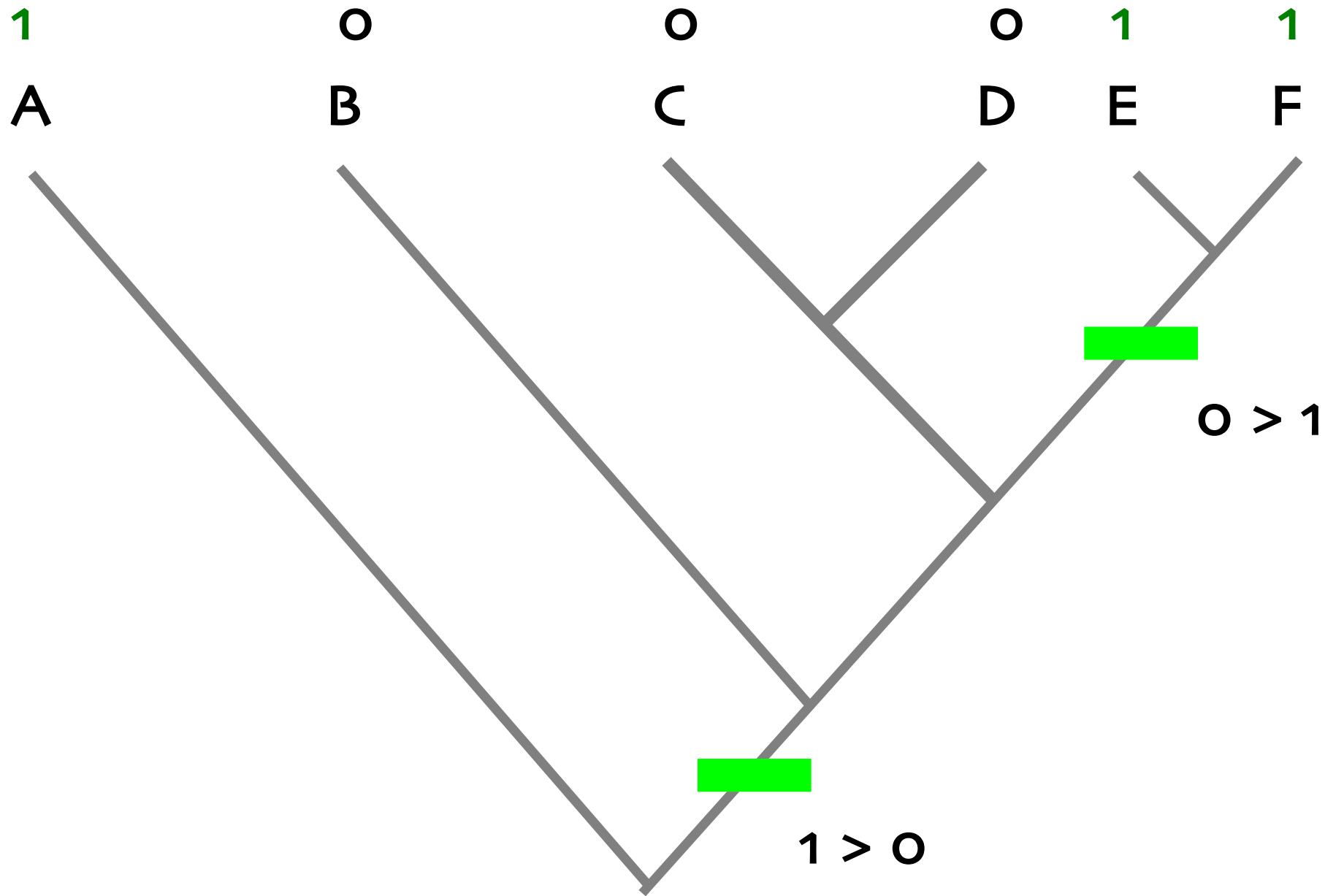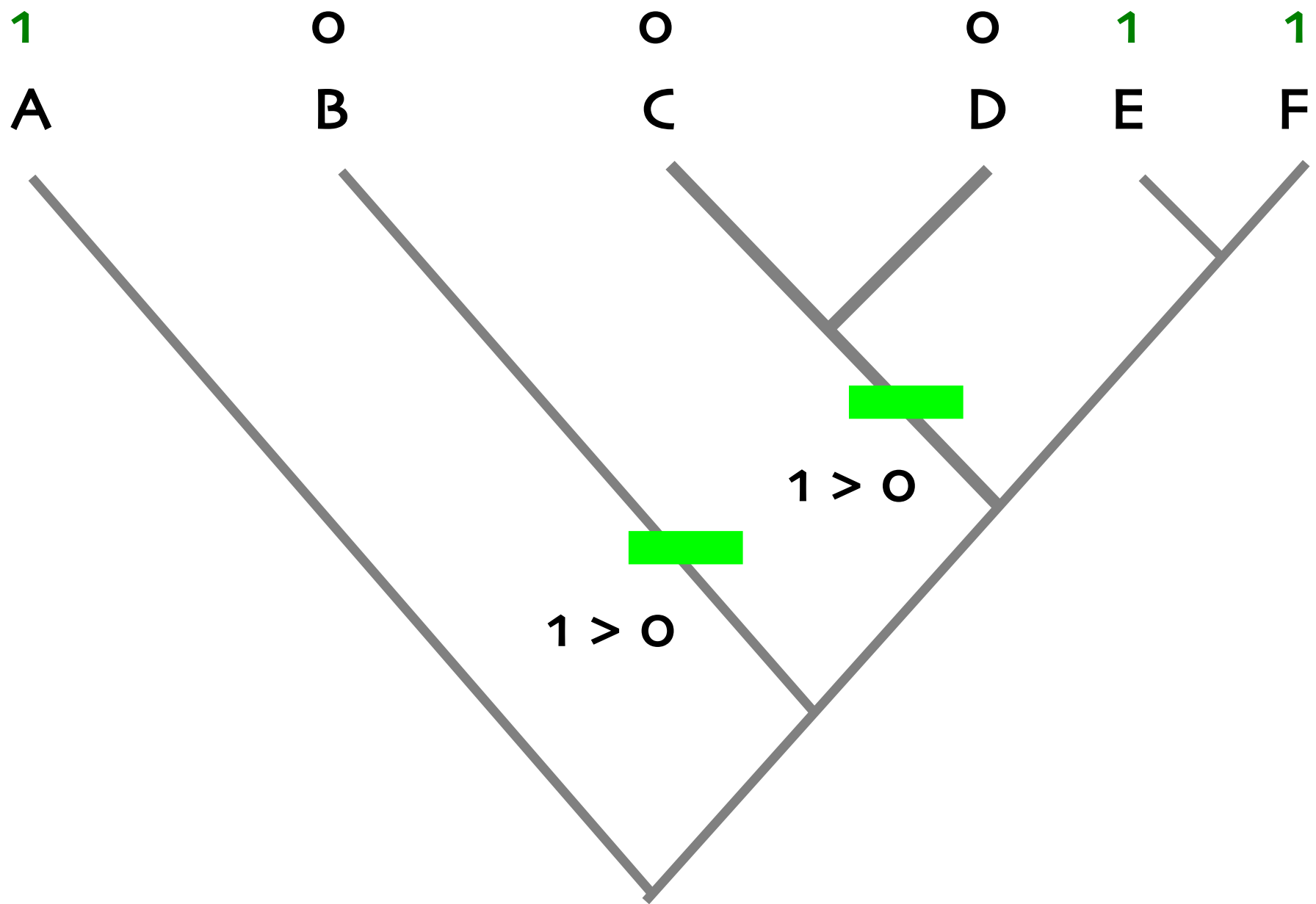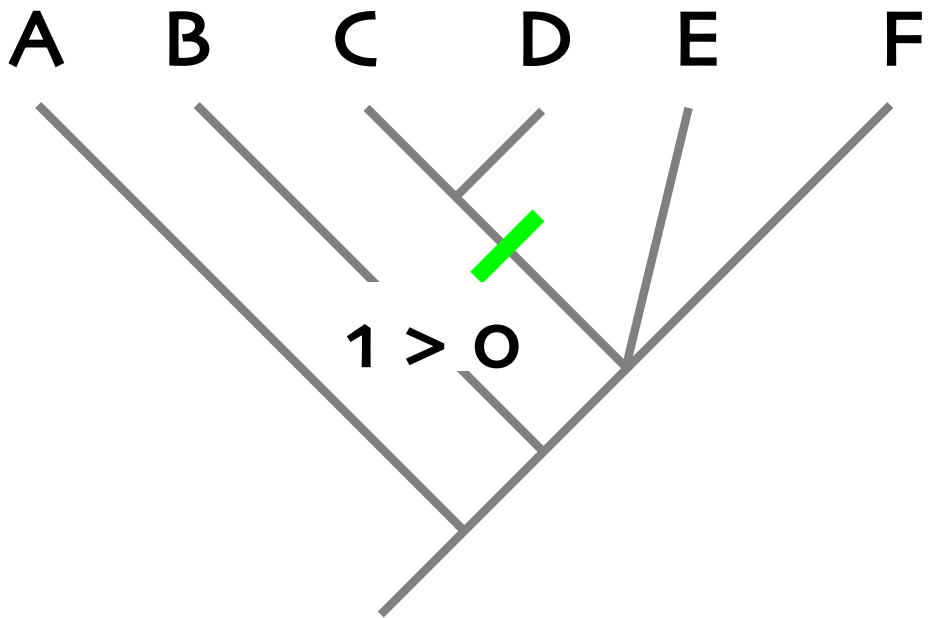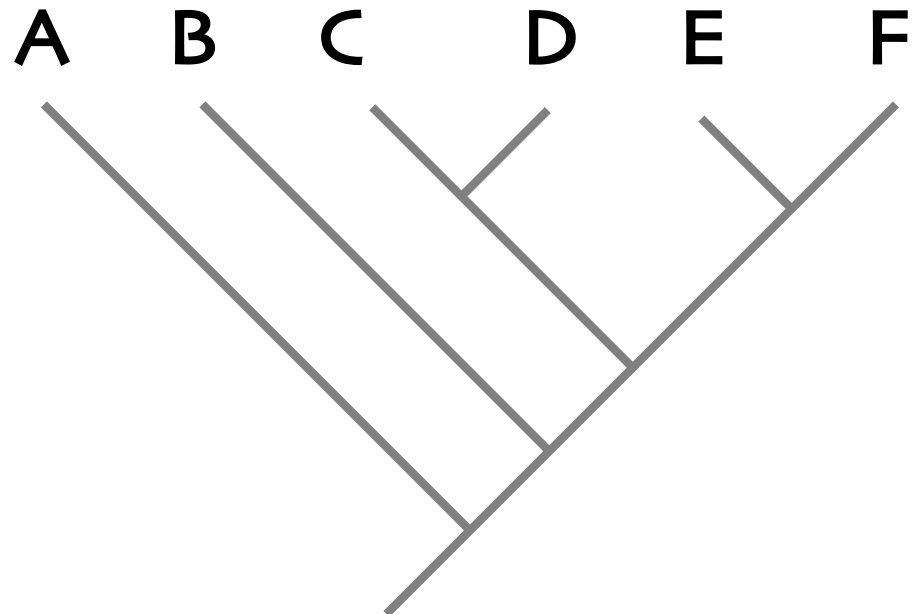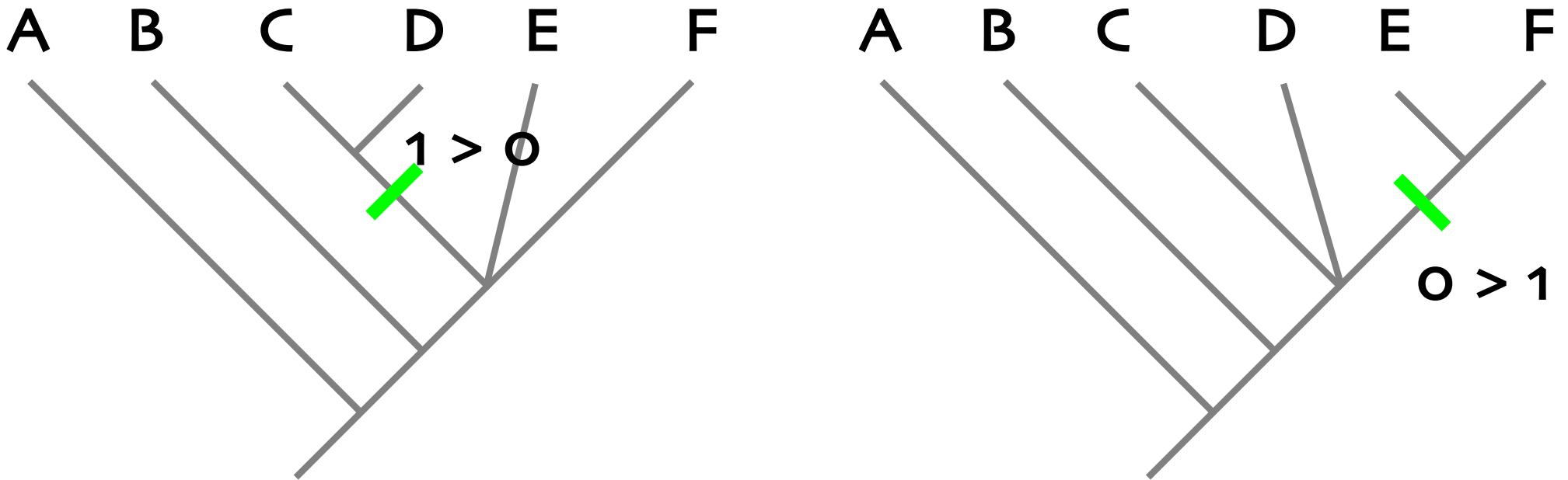
both changes **impossible** <u>simultaneously</u>

A B C D E F
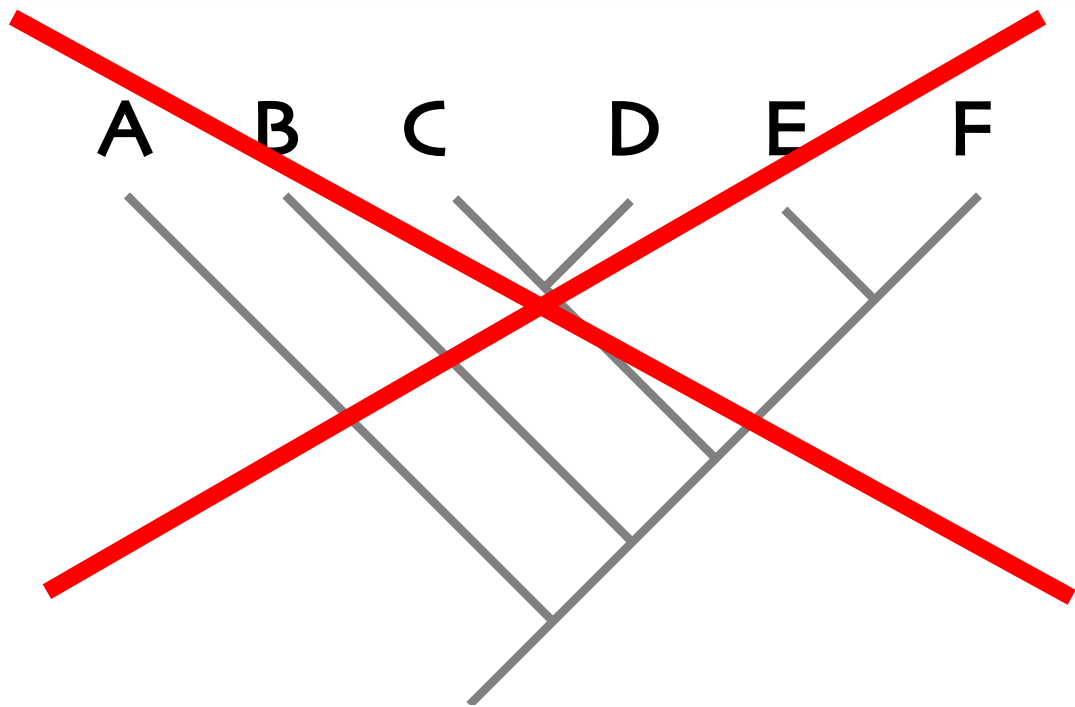
"Super"trees

original optimal TREES combined

aim to combine results of **_SEPARATE_** analyses

unlike in previous approaches trees with **NON MATCHING** terminals might be combined

cut & glue
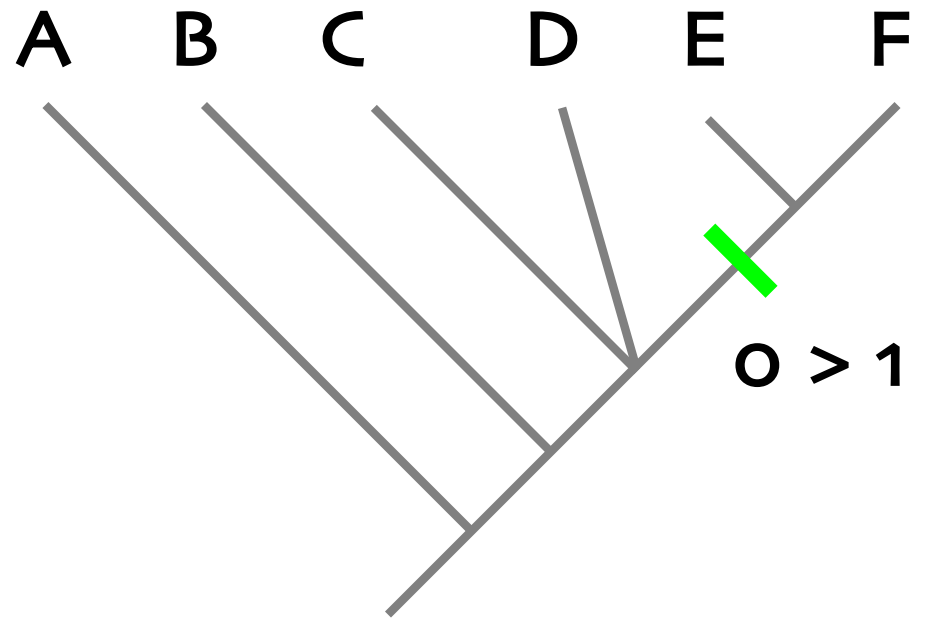
aiming for synthesis without proper analysis

fast & easy

originally presented for handling LARGE materials because they were considered as impossible to analyze

"super"trees can include groups ABSENT from ALL original trees & lack those present in ALL original trees

Goloboff, P. A. 2005. Minority rule supertrees? MRP, compatibility, and minimum flip may display the _least_ frequent groups. _Cladistics_ 21: 282-294.

"Super"trees suck

SIMULTANEOUS ANALYSIS

```
  0000000000111111111122222222223333333333444444444455555555556
  1234567890123456789012345678901234567890123456789012345678901234567890
```

	pollen	seedling	vegetative parts	flower/inflorescense	fruit
A	0100011100	0101011010	1000111010	0011101111110001	11111110111101
B	0111110000	0100100100	1011001000	1101111111110110	011001010001000
C	0101101001	0001111010	0011001111	1111000111011000	001111111001010
D	1011101100	0001110101	0011000101	0011110000000011	0111001000010
E	0100101000	1101111000	0100001000	1000101001000001	0110110111101
F	1000111001	0101100010	1011100110	1010110011110110	1001110101100

pollen   seedling   vegetative parts   flower/inflorescense   fruit

SIMULTANEOUS ANALYSIS

```
    00000000011111111112222222222333333333344444444445555555556
    12345678901234567890123456789012345678901234567890123456789 0
```

A	010001110	010101101010 0	0111010001	1011111110001	11111110111101
B	011111000	010010010010 1	1001000110	1111111101100	11001010001000
C	010110100	100011101001 1	1001111111	1000111011000 0	01111111001010
D	101110110	000011101010 0	1100010100	1111000000011	01111001000010
E	010010100	011011110000 1	0000100010	0010100100000 1	01101101111101
F	100011100	101011000101 0	1110011010	1011001111011 0	11001110101100

spores　　　thallus　　　asci　　　secondary chemistry　　　DNA

SIMULTANEOUS ANALYSIS

| matrix 1 |
| matrix 2 |
| matrix 3 |
| matrix 4 |
| matrix 5 |

separate analyses

| tree 1 |
| tree |
| tree 3 |
| tree |
| tree 5 |

consensus approach

trees combined

result

on final tree only clades common to ALL trees

| matrix 1 |
| matrix 2 |
| matrix 3 |
| matrix 4 |
| matrix 5 |

matrices combined

combined matrix

analysis

result

simultaneous analysis
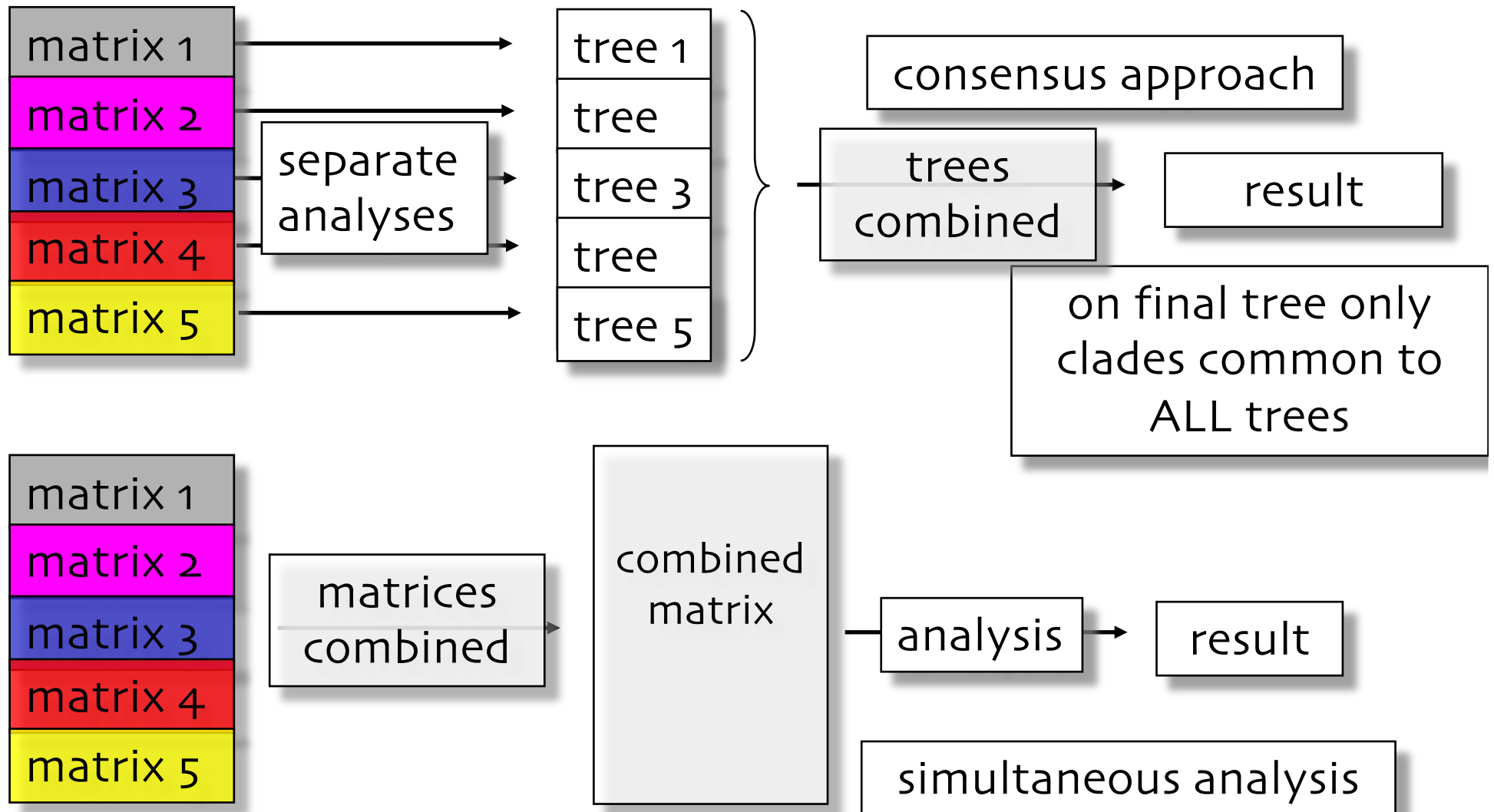
SIMULTANEOUS ANALYSIS
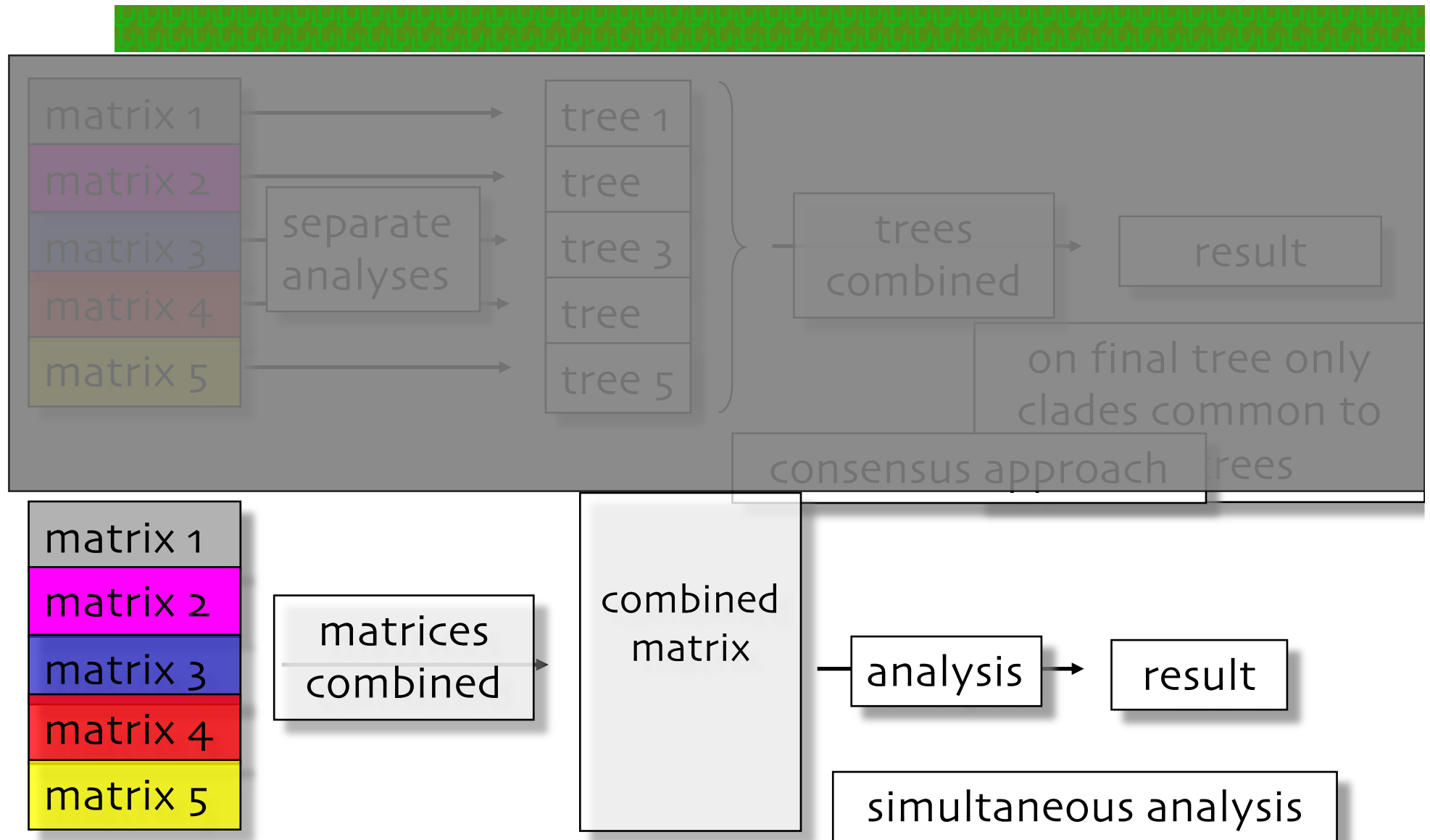
CHARACTER CONGRUENCE

is the concrete reason for the fact that we find short tree(s)

on this kind of tree evolutionary changes are
CONGRUENT, i.e. CONCENTRATED ON SAME
BRANCHES of the tree (less homoplasy)

1. Is this accidental?

2. Or have we possibly find a hypothesis that is a good
approximate about evolutionary history and this is
why so MANY characters are congruent?

DESCENT WITH MODIFICATION

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

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

SIMULTANEOUS ANALYSIS

correlation between homoplastic characters that would overcome signal based on **homologies** is <u>extremely unlikely</u> (practically impossible) especially when characters included are from different levels (from nucleotides to macromorphology), represent different phases of life-cycle etc.

..foibles of the individual sources of information are not generally shared. Since ubiquitin follows one set of rules and 18S another, the only source of shared information is **history**…

Wheeler, W. C. ym. 1993. Arthropod phylogeny: combined approach. *Cladistics* 9:1-39.

SIMULTANEOUS ANALYSIS

signal obtained from individual
matrices might be weak and
concealed by homoplasy
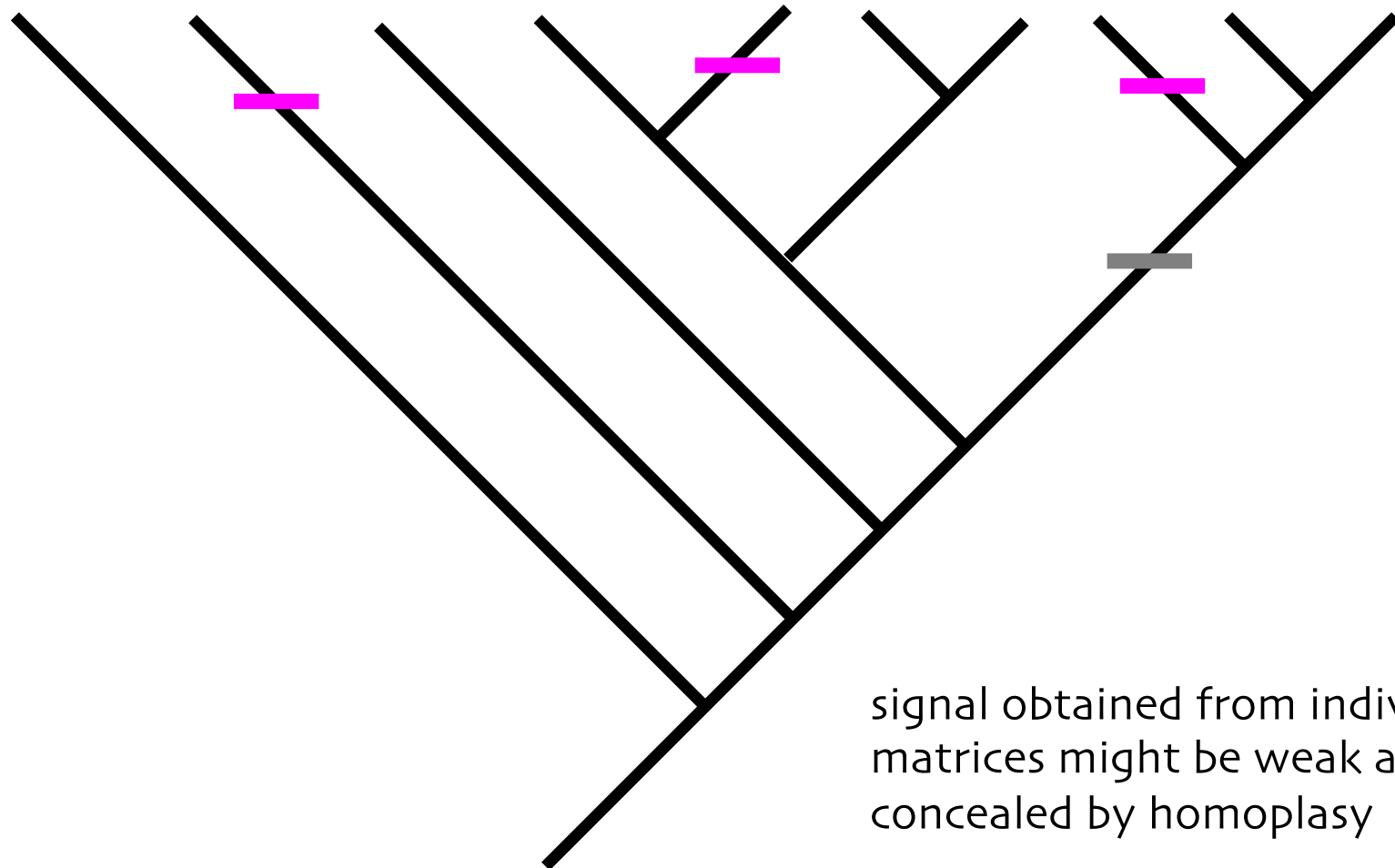
SIMULTANEOUS ANALYSIS

combination of ALL matrices will AUTOMATICALLY reveal signal of evolutionary history if such exists

character congruence

.... or we can even solve problematic parts of the phylogeny that have remained ambiguous before

16.xi.

1. compromise trees (continued...)

2. simultaneous analysis

3. search algorithms

4. traditional search strategies

5. new search strategies

6. summary

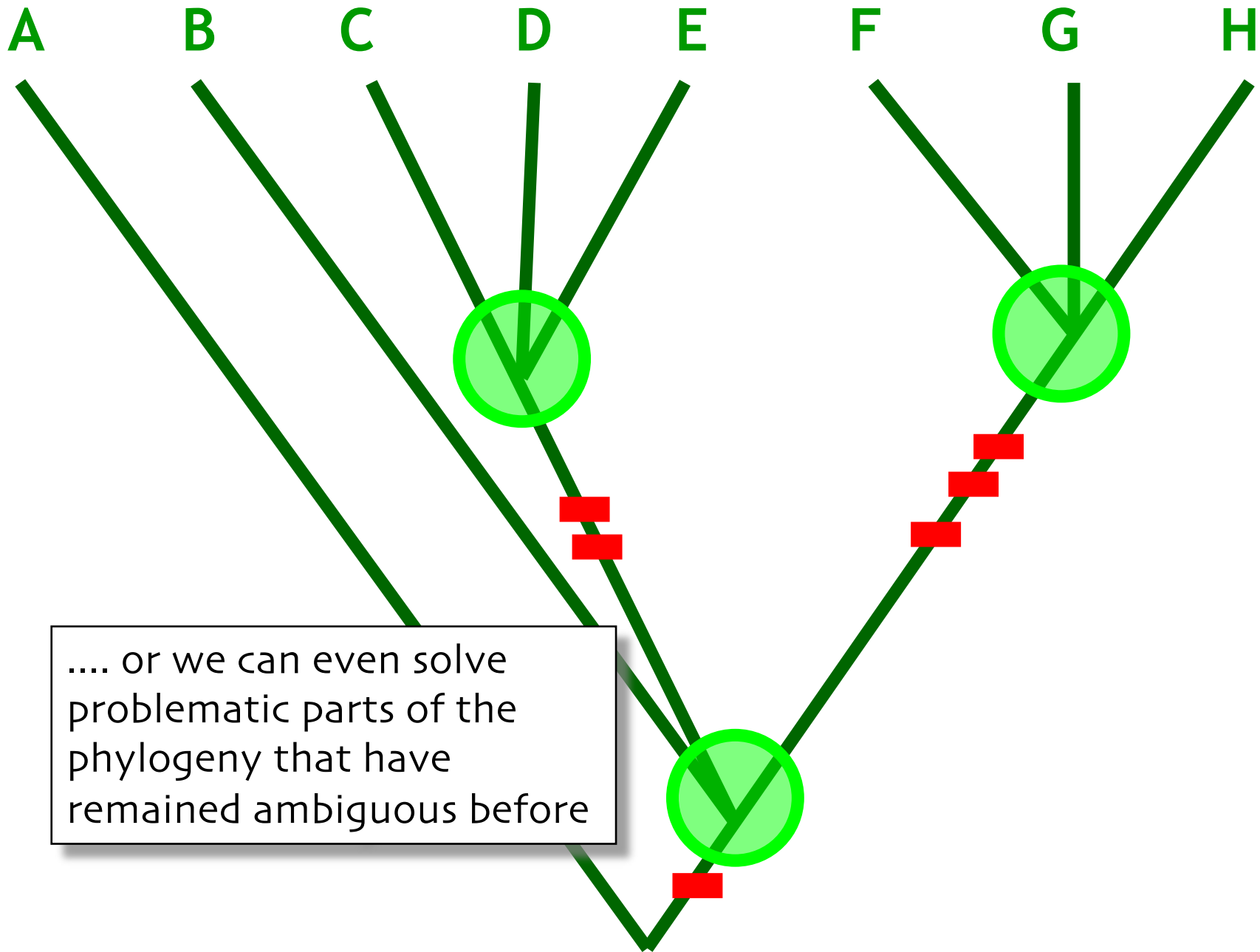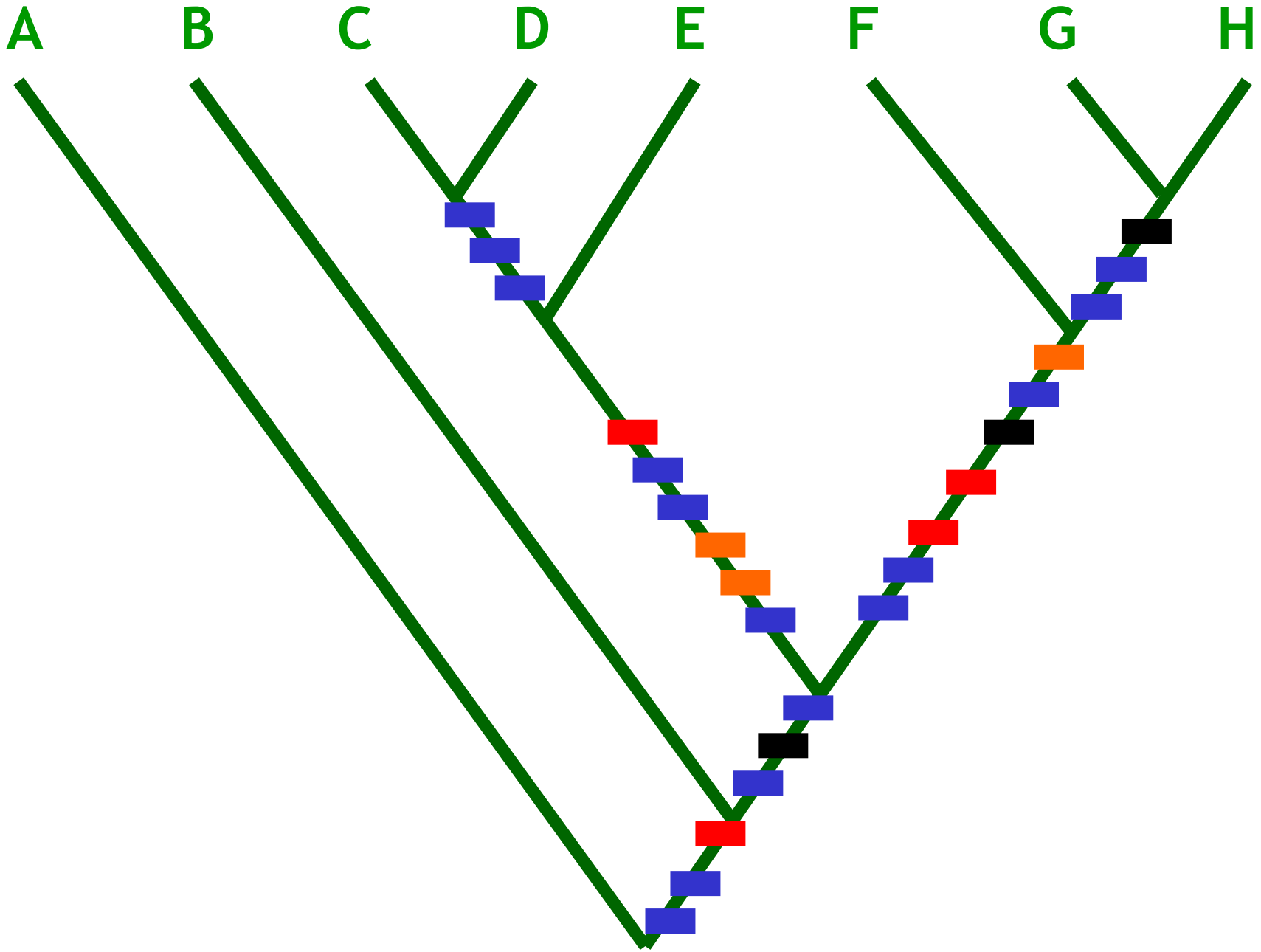| trees	0 1	0 2	0 3	0 4	0 5	0 6	0 7	0 8	0 9	1 0	1 1	1 2	Σ
1	1	1	2	2	2	2	2	1	2	2	1	2	20
2	1	1	2	2	2	2	2	1	2	2	2	2	21
3	1	1	2	2	2	2	2	1	1	2	2	2	20
4	1	2	2	1	2	2	2	1	2	2	2	2	21
5	1	2	2	1	2	2	2	1	2	2	2	1	20
6	1	2	2	1	2	2	2	1	2	2	2	2	21
7	1	2	2	2	2	2	2	1	2	1	1	2	20
8	1	2	1	2	2	1	2	1	2	1	2	2	19
9	1	2	2	2	2	2	2	1	2	1	2	1	20
10	1	2	2	2	2	2	2	1	1	2	2	1	20
11	1	2	2	2	2	2	2	1	2	2	1	2	21
12	1	2	1	2	2	1	2	1	2	2	2	2	20
13	1	2	2	2	1	2	1	1	1	2	2	2	19
14	1	2	2	2	1	2	1	1	2	2	2	2	20
15	1	2	1	2	1	1	1	1	2	2	2	2	18

characters

T	B(T)
3	1
4	3
5	15
6	105
7	945
8	10 395
9	135 135
10	2×10^{6}
15	8×10^{12}
20	2×10^{20}
50	3×10^{74}

exhaustive search possible for only limited no. of terminals
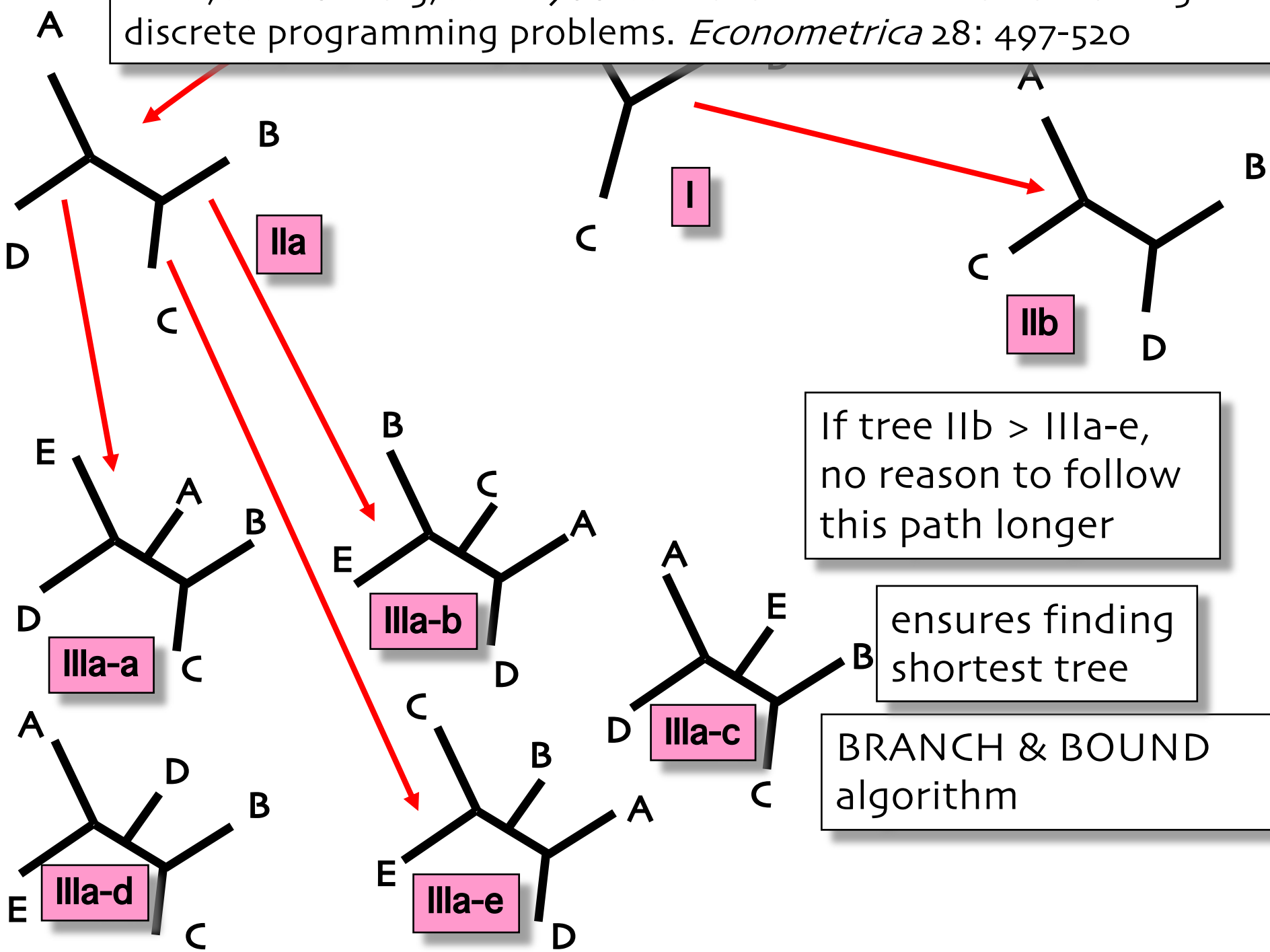
Land, A.H. & Doig, A.G. 1960. An automated method of solving discrete programming problems. *Econometrica* 28: 497-520

A

B

D

C

IIa

I

C

A

B

C

IIb

D

E

A

B

D

C

IIIa-a

B

C

E

A

IIIa-b

D

If tree IIb > IIIa-e, no reason to follow this path longer

A

E

B

D

C

IIIa-c

ensures finding shortest tree

BRANCH & BOUND algorithm

A

D

B

E

IIIa-d

C

C

B

A

E

IIIa-e

D

T	B(T)
3	1
	3
	15
6	105
7	945
8	10 395
9	135 135
10	2×10^6
15	8×10^{12}
20	2×10^{20}
50	3×10^{74}

1,7 GHz processor

$1,65 \times 10^6$ trees/second

BRANCH and BOUND algorithm

can be used for max. ca. 30 terminals

TSP 🌐

www.math.uwaterloo.ca/tsp/history/milestone.html

Milestones in the Solution of TSP Instances

Computer codes for the TSP have become increasingly more sophisticated over the years. A conspicuous sign of these improvements is the increasing size of nontrivial instances that have been solved, moving from Dantzig, Fulkerson, and Johnson's solution of a 49-city problem in 1954 up through the solution of a 24,978-city problem 50 years later.

Year	Research Team	Size of Instance	Name
1954	G. Dantzig, R. Fulkerson, and S. Johnson	49 cities	dantzig42
1971	M. Held and R.M. Karp	64 cities	64 random points
1975	P.M. Camerini, L. Fratta, and F. Maffioli	67 cities	67 random points
1977	M. Grötschel	120 cities	gr120
1980	H. Crowder and M.W. Padberg	318 cities	lin318
1987	M. Padberg and G. Rinaldi	532 cities	att532
1987	M. Grötschel and O. Holland	666 cities	gr666
1987	M. Padberg and G. Rinaldi	2,392 cities	pr2392
1994	D. Applegate, R. Bixby, V. Chvátal, and W. Cook	7,397 cities	pla7397
1998	D. Applegate, R. Bixby, V. Chvátal, and W. Cook	13,509 cities	usa13509
2001	D. Applegate, R. Bixby, V. Chvátal, and W. Cook	15,112 cities	d15112
2004	D. Applegate, R. Bixby, V. Chvátal, W. Cook, and K. Helsgaun	24,978 cities	sw24798

TSP

Sweden Tour

24,978 Cities

0   50   100 Kilometers
0   50   100 Miles

Lambert Conformal Conic Projection, SP 47N/62N

72 500 km

Foulds, L.R. & Graham, R.L. 1982. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics* 3: 43–49.

Traveling salesman retires

NP complete/hard

X. Xxxxxxxxx, X. Xxxxxxx & X. Xxxxxxx

Xxxxxx, PO Box XX, XX-XXXXX University of XXXX, XXXXXX

THE search of all equally parsimonious trees has formerly been described as a NP complete problem without possibility to find all possible trees. We present a new way to ensure that all these trees will be found even during the heuristic search by estimation of the

HEURISTIC SEARCH

Wagner algorithm

Kluge, A. G. & Farris, J. S. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Zoology* 18:1-32.
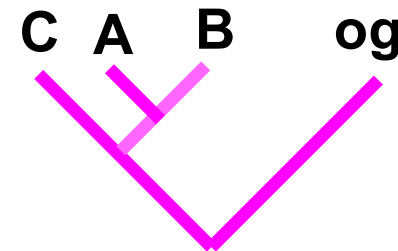
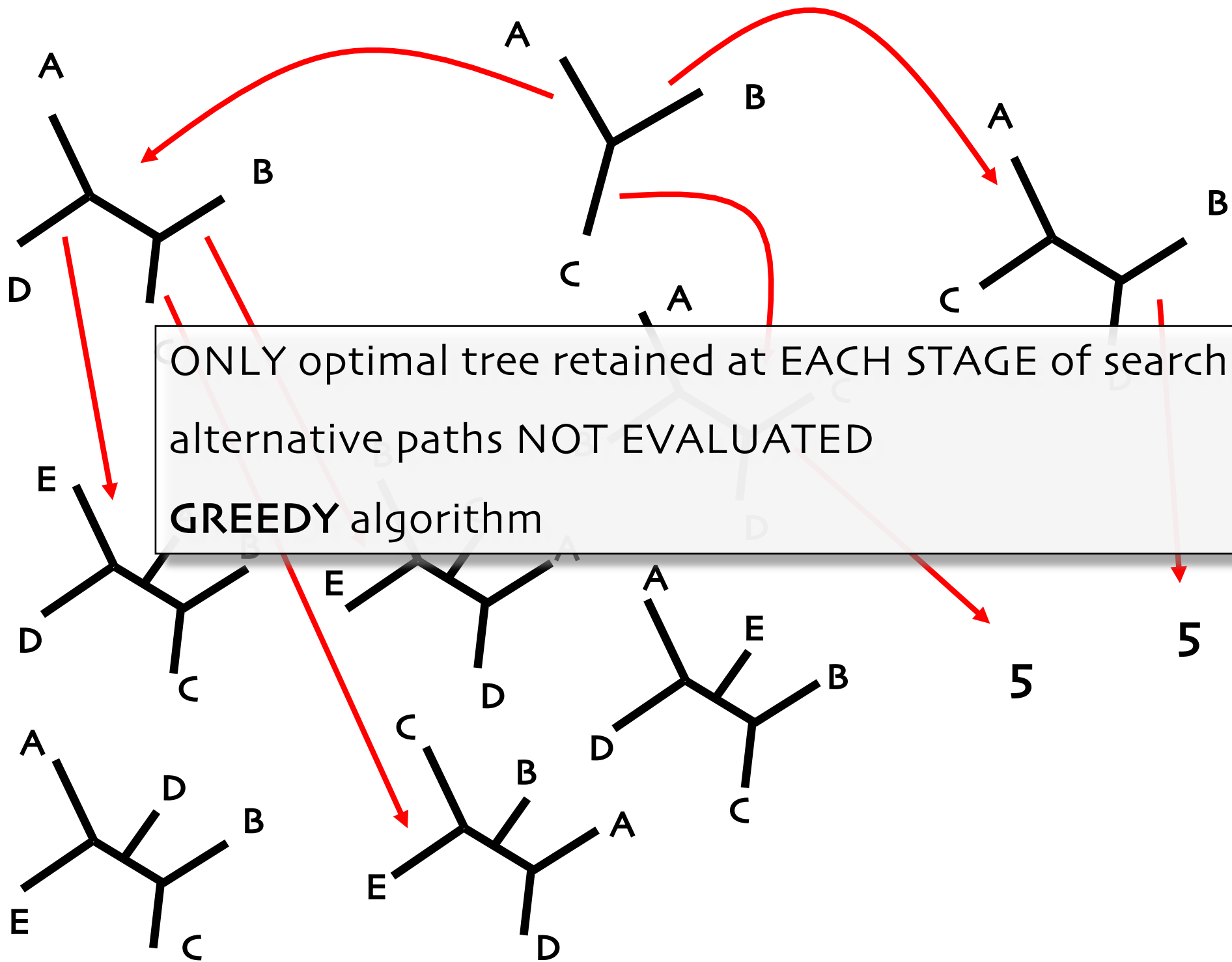Farris, J. S. 1970. Methods for computing Wagner trees. *Systematic Zoology* 19:83-92.

	1	2	3	4	5	6	7	8	9	10	differences from outgroup
og	0	0	0	0	0	0	0	0	0	0	-
A	1	0	0	0	1	1	0	0	0	1	4
B	1	0	0	1	0	0	0	0	0	1	3
C	0	0	0	0	0	0	0	0	1	1	2
D	0	1	1	0	0	0	1	1	0	1	5
E	0	1	1	1	0	0	0	1	0	1	5

the problem of Wagner algorithm is that the order where terminals are added to tree affects the tree finally obtained

CONCRETE reason for this?

C A B og

continued until all terminals are included

ONLY optimal tree retained at EACH STAGE of search

alternative paths NOT EVALUATED

GREEDY algorithm

HEURISTIC SEARCH

RAS = random addition sequence

1. starting trees build by adding terminals one by one

matrix/random order

Wagner algorithm

phenetic clustering

all these can be combined to build starting trees
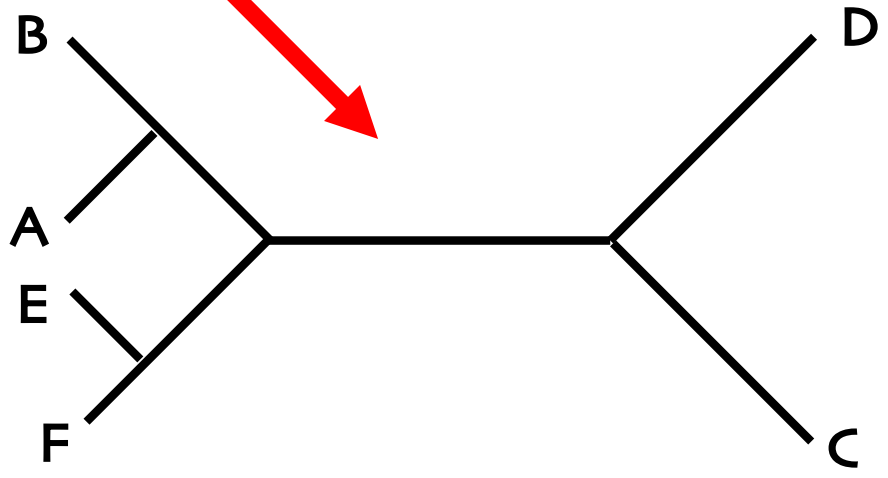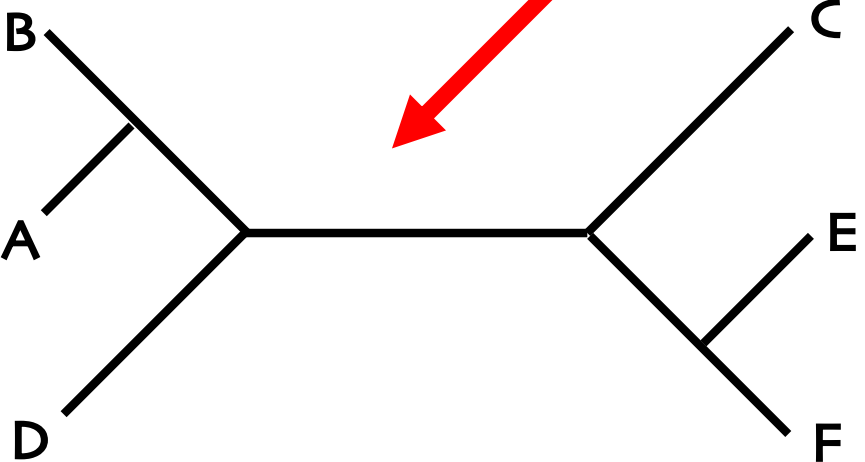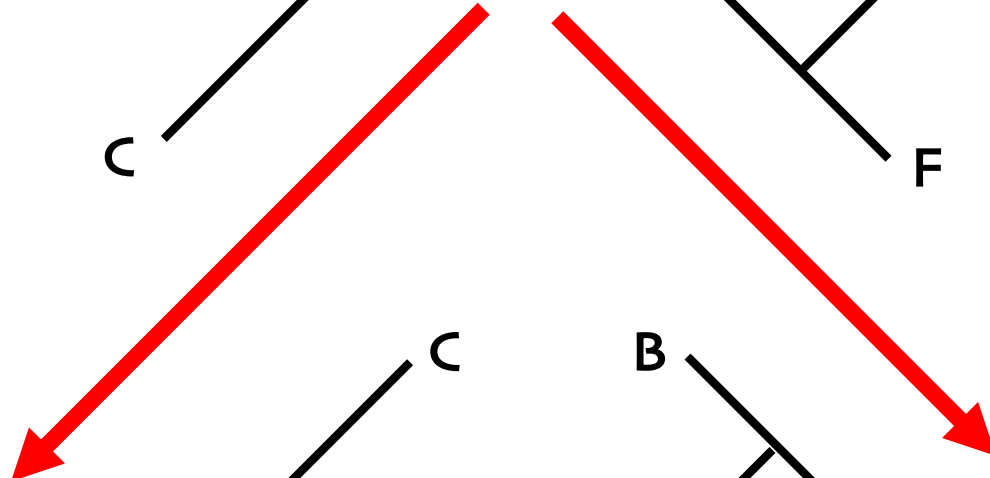
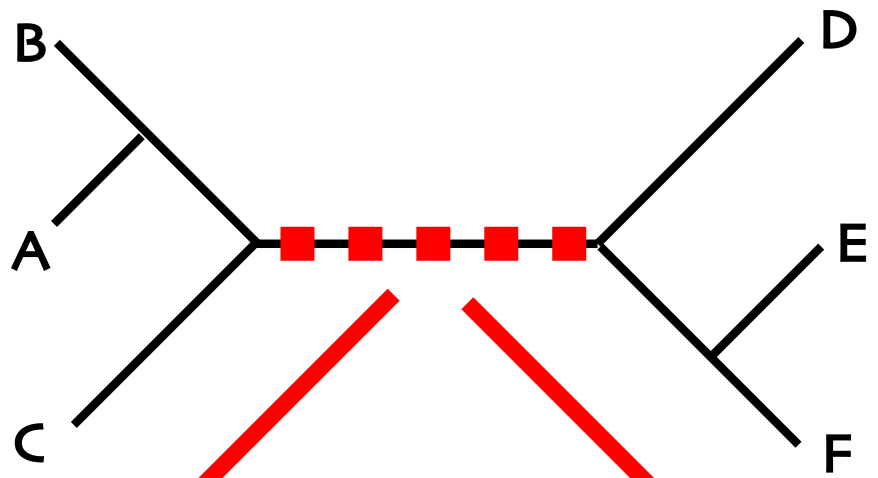leads RARELY in finding shortest tree

HEURISTIC SEARCH

1. starting trees build by adding terminals one by one

2. after this branches of the tree(s) are moved

branch SWAPPING

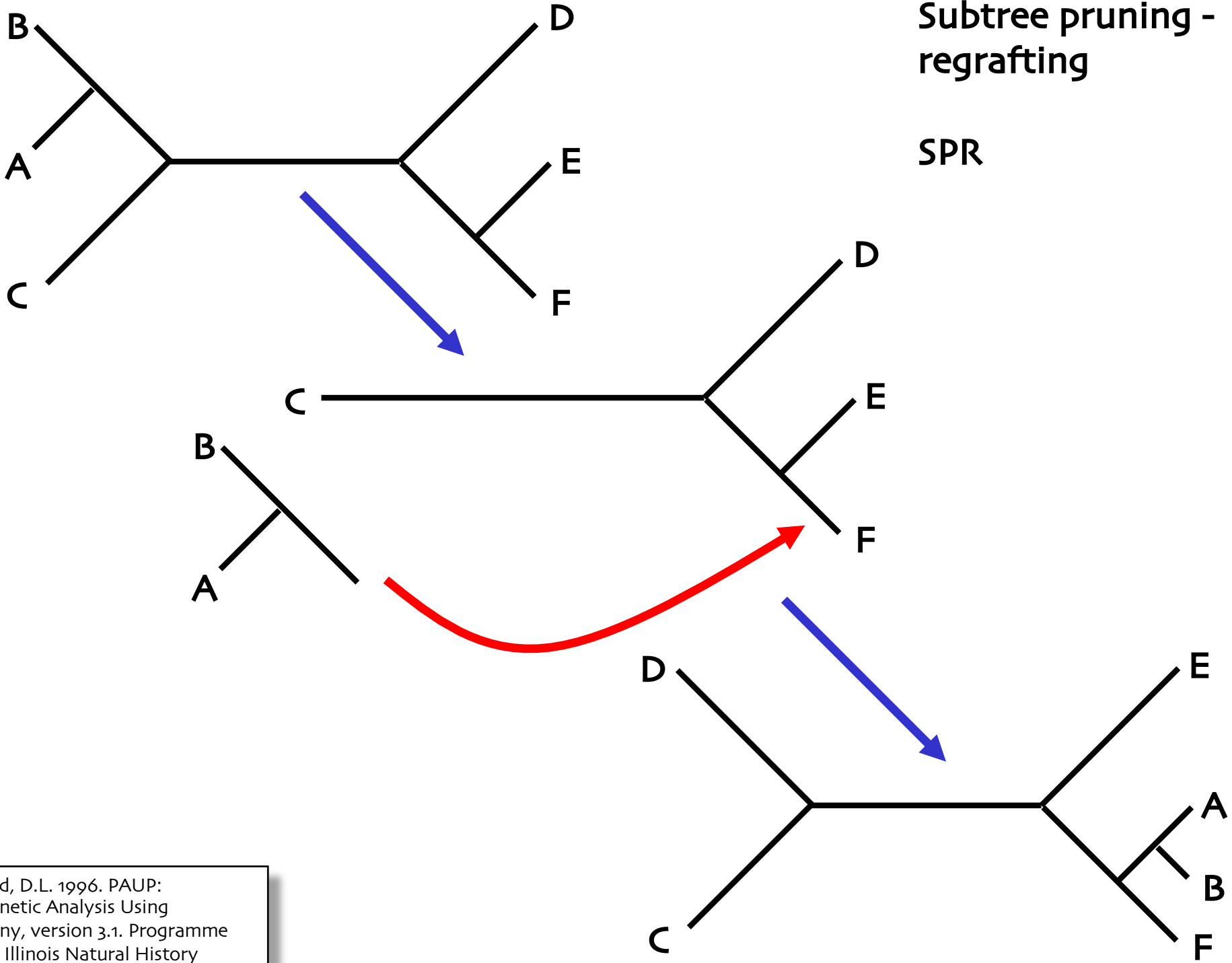**Nearest-neighbor interchange**

NNI

Camin & Sokal 1965

Swofford, D.L. 1996. PAUP: Phylogenetic Analysis Using Parsimony, version 3.1. Programme manual. Illinois Natural History Survey.
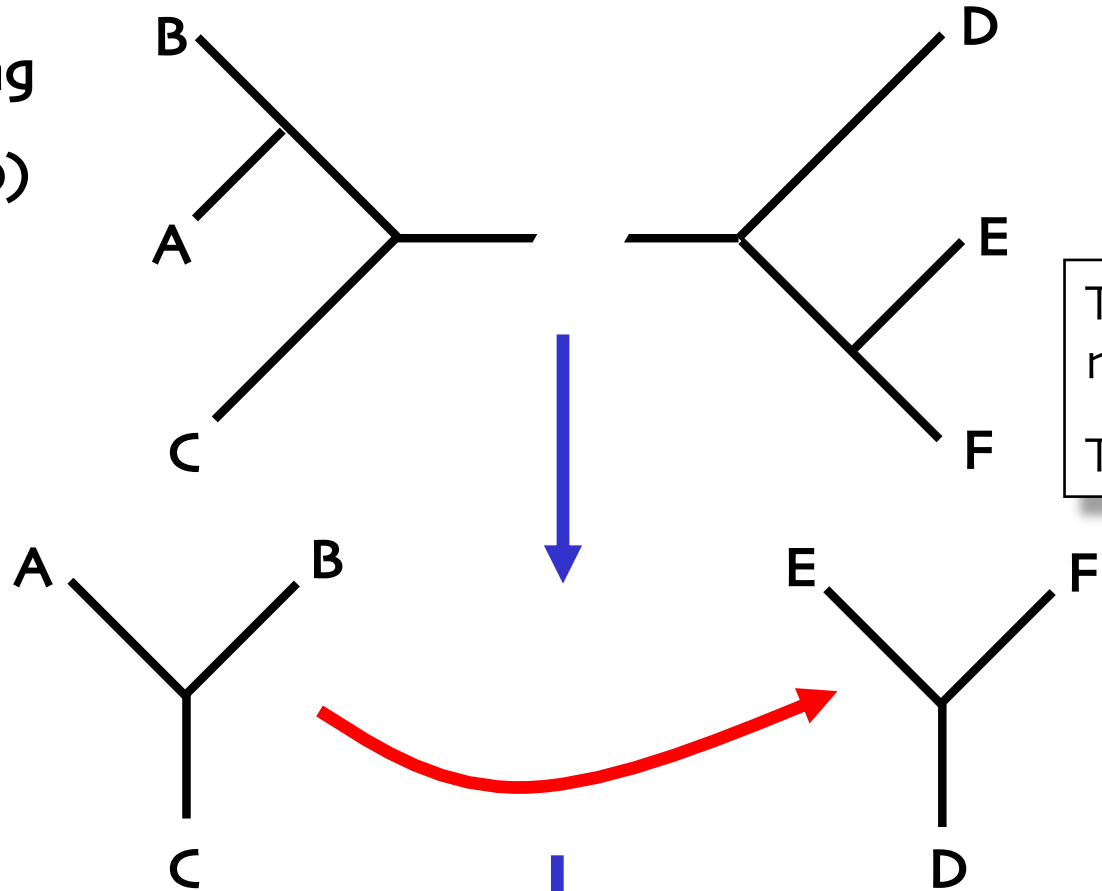
Subtree pruning - regrafting

SPR

Swofford, D.L. 1996. PAUP: Phylogenetic Analysis Using Parsimony, version 3.1. Programme manual. Illinois Natural History Survey.
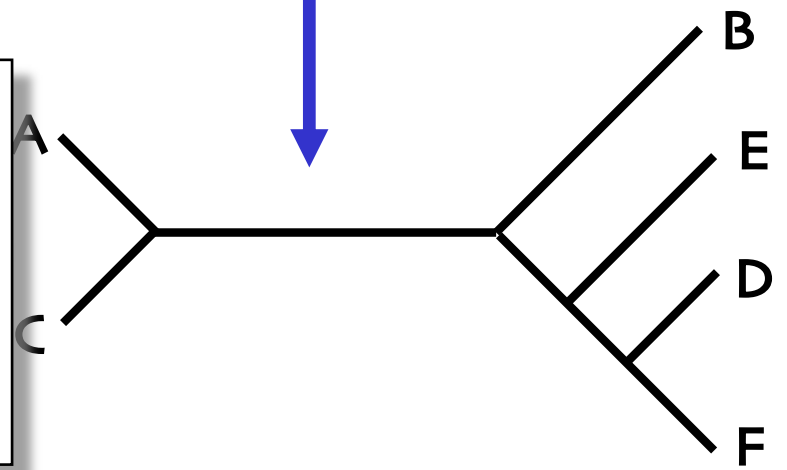
Branch breaking

BB (Farris 1970)

Tree bisection-reconnection

TBR (Swofford 1990)

for small matrices (< 50 spp.) surprisingly efficient

despite of the fact that only an extremely SMALL FRACTION of all possible trees considered

(50 spp. 1.0×10^{-69})

Swofford, D.L. 1996. PAUP: Phylogenetic Analysis Using Parsimony, version 3.1. Programme manual. Illinois Natural History Survey.

HEURISTIC SEARCH

1. starting trees build by adding terminals one by one

2. after this branches of the tree(s) are moved

   branch SWAPPING with NNI, SPR, TBR

3. current programs include algorithms that allow
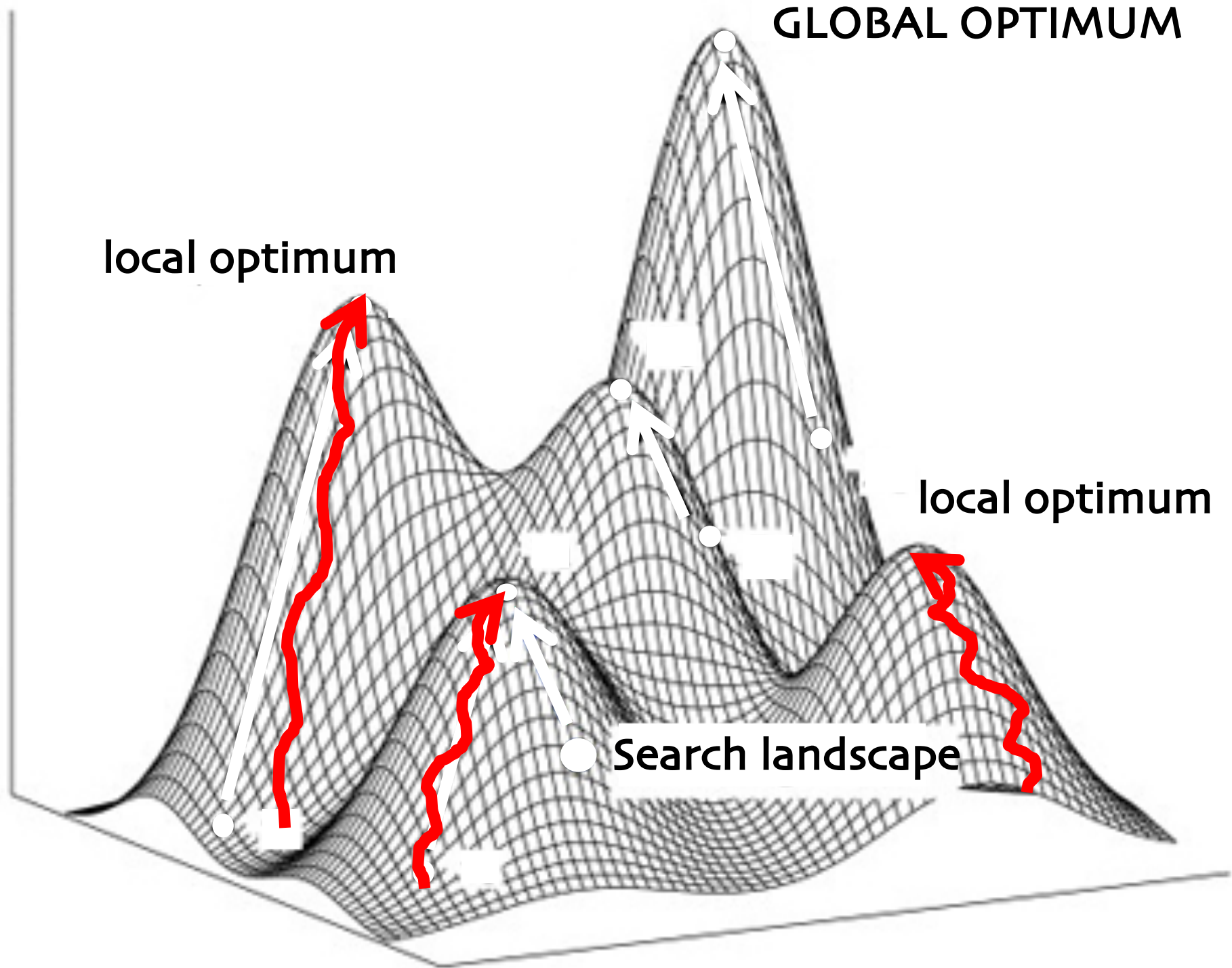   evaluation of tree length WITHOUT visiting ALL
   nodes of trees

   save processing time

Goloboff, P.A. 1993. Character optimization and
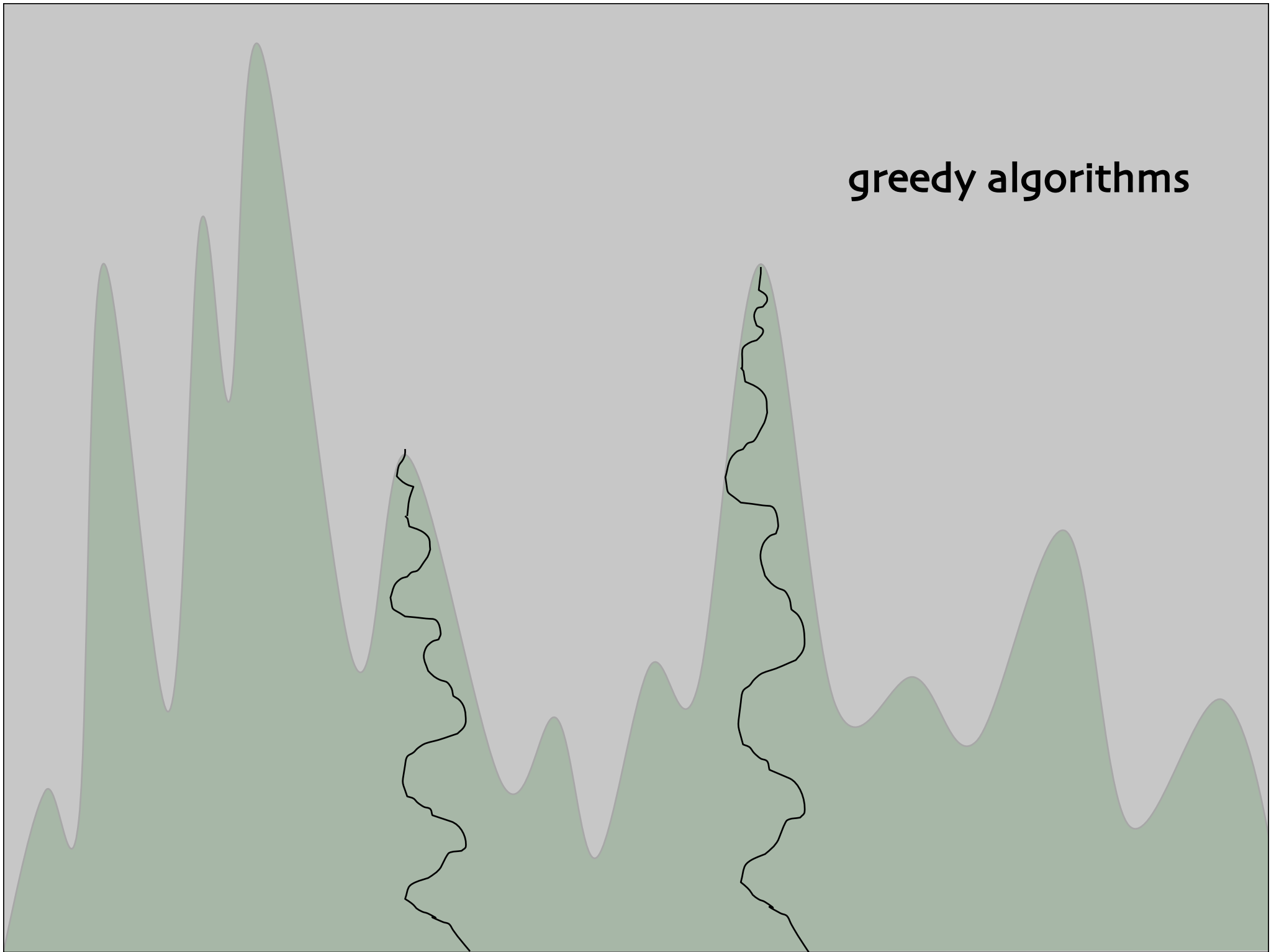calculation of tree lengths. *Cladistics* 9: 433-436.
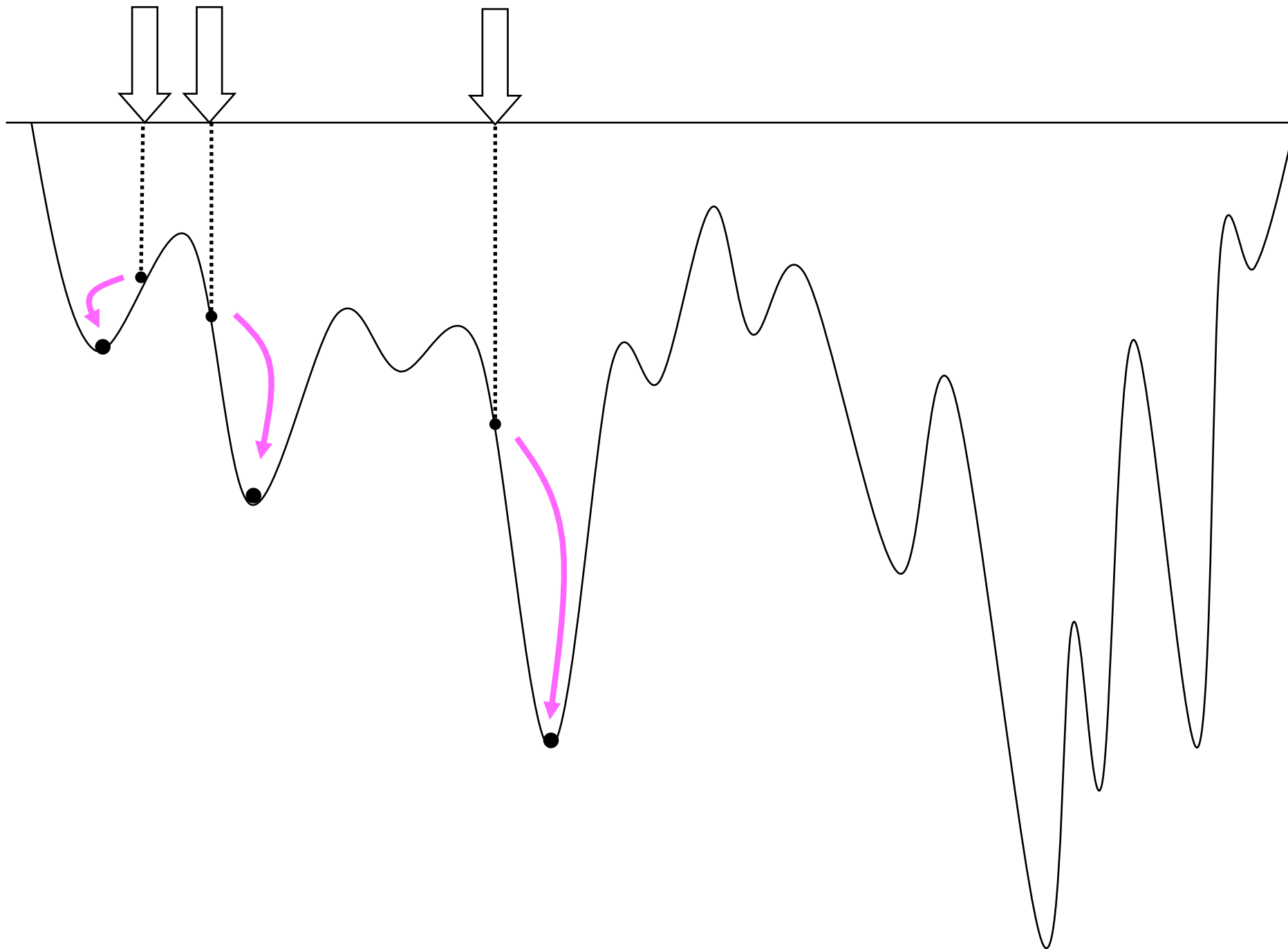
Search strategies

islands of trees in tree "space" (landscape)

problem especially in analyses of
larger matrices (> 100-200 terminals)

GLOBAL OPTIMUM

local optimum

local optimum

Search landscape

greedy algorithms

GREEDY algorithms simple, do NOT retrace

example 2 dimensional real tree space
MULTI dimensional

best way to move around in treespace?

T	B(T)
10	2×10^{6}
15	8×10^{12}
20	2×10^{20}
50	3×10^{74}
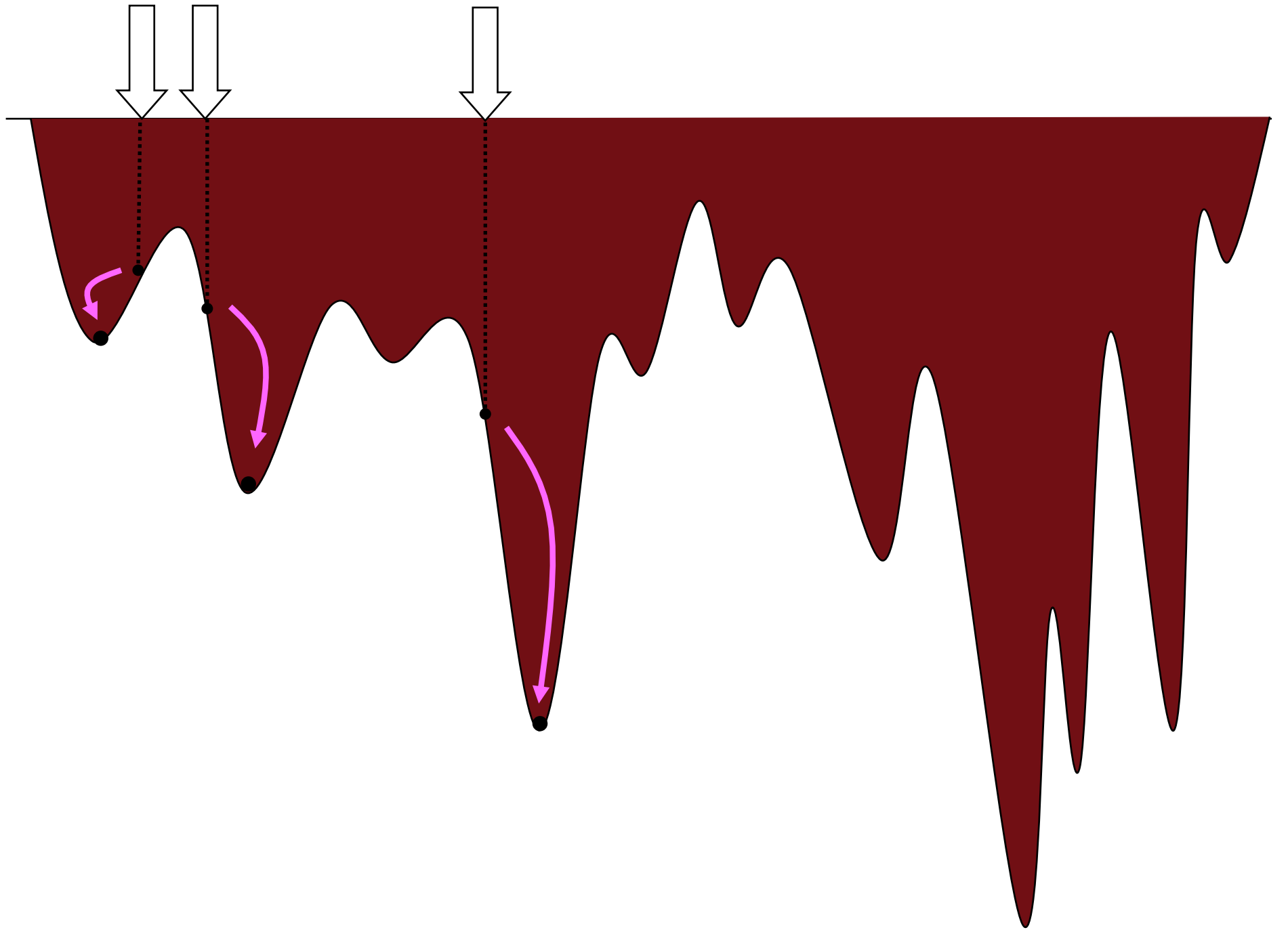
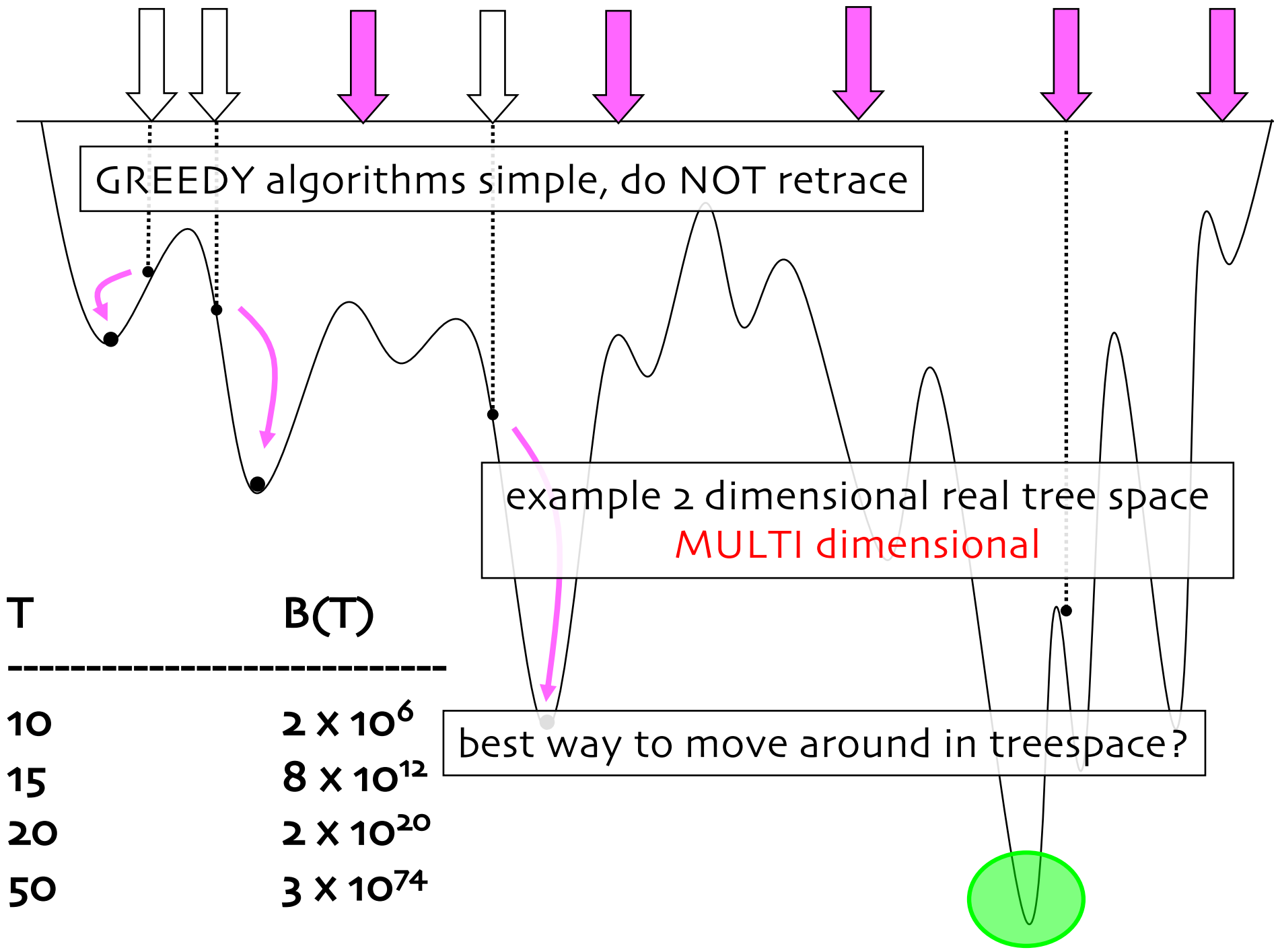# TRADITIONAL SEARCH STRATEGIES

search repeated SEVERAL TIMES (10 - 10 000 x)

how many times is sufficient? | has optimal tree be found? |

if first 100x give same result | STOP |

if only 1/100 give optimal result | CONTINUE |

moving branches takes time, thus the number of trees
    saved at intermediate stages of search kept
    SMALL, e.g. $\leq$ 2-10 | WHY? |

when starting search potentially VERY DIFFERENT
    trees, after moving branches this is NOT
    anymore true, no reason to keep MANY
    SIMILAR trees | most likely will lead finding same optimal tree |

TRADITIONAL SEARCH STRATEGIES

if > 10% of random searches give same result --->
search should be extended by expanding the
number of trees saved in cache memory (50-100)

if < 10% random searches give same results --->
continue with NUMEROUS (500-1 000) random search
with saving only 2-10 trees in cache

5-10% of trees saved in cache (50-200)
   analyzed in more detail

Davis, J. & al. 2005. The limits of conventional cladistic analysis. In: Albert, V. (ed.)
Parsimony, phylogeny and genomics. 229 s. Oxford University Press

TRADITIONAL SEARCH STRATEGIES

NO NEED to find ALL optimal trees (Farris & al. 1996)

for large matrices > 10^6 equally parsimonious trees
    might exist

consensus tree based only on small sample of trees
might be identical with the one based on ALL trees if
those sampled originate from SEPARATE tree islands

can also be used to "guide" our search, i.e. if consensus
does not change, no need to continue

NEW SEARCH STRATEGIES

required for analyses of LARGE (> 500-700 spp.) matrices

in large trees ALL parts should be in optimal arrangement
in order to be optimal as a WHOLE
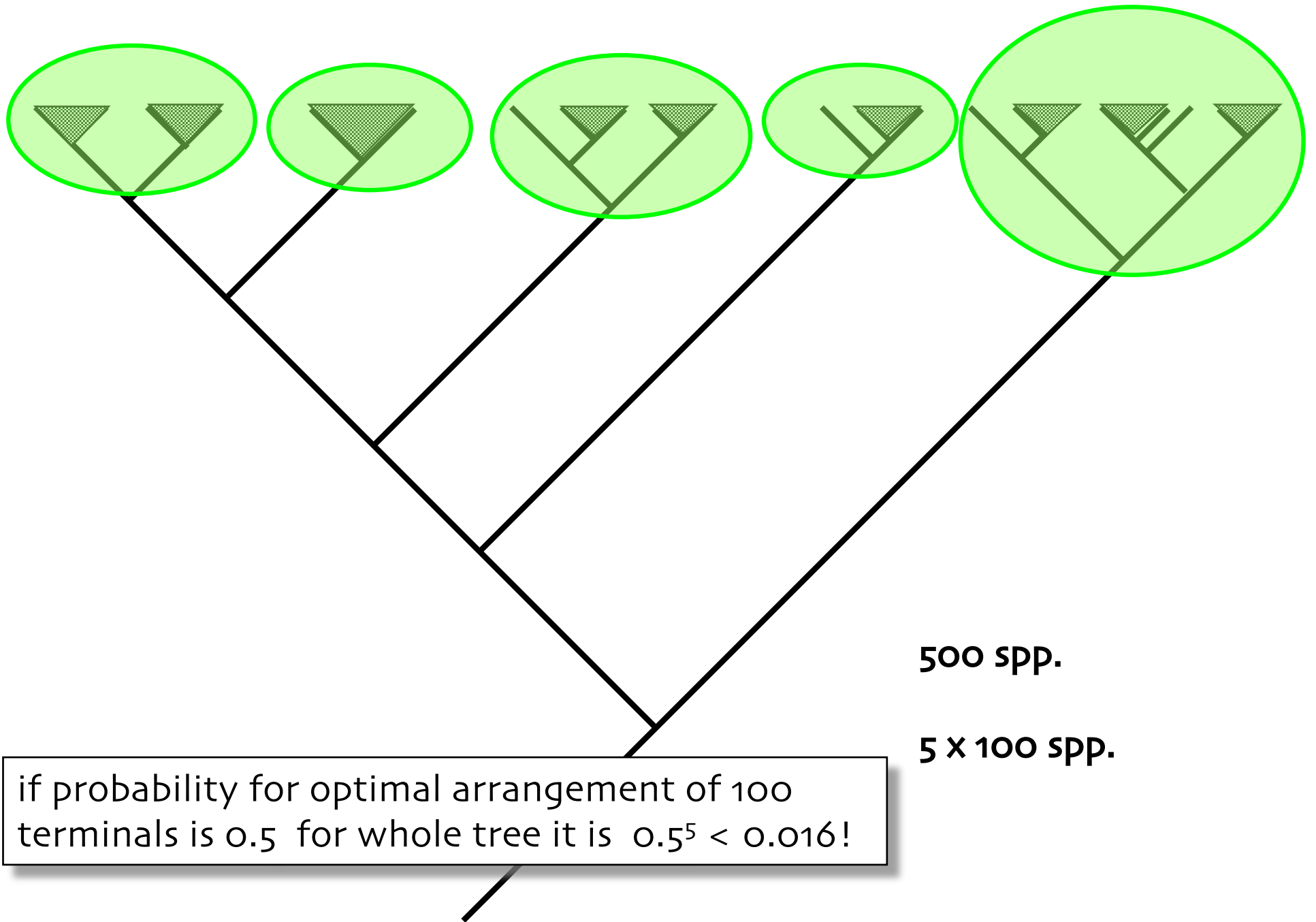
500 spp.

5 x 100 spp.

if probability for optimal arrangement of 100 terminals is 0.5 for whole tree it is $0.5^5 < 0.016$!

GOLOBOFF 2000: dice comparison

www.weld-re1.k12.co.us/webclass/students/reamj/web/dice.gif

NEW SEARCH STRATEGIES

Moilanen, A. 1999. Searching for most parsimonious trees with simulated evolutionary optimization. *Cladistics* 15: 39-50.

Nixon, K.C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15: 407-414.

Goloboff. P.A. 1999. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* 15: 415-428.

Nixon, K. Parsimony ratchet

1a. starting tree (e.g. Wagner algorithm)
1b. continued using fast & simple branch swapping

2. weight randomly e.g. 10-30% of characters
3. try to find shortest tree of this NEW matrix (as in 1b)

4. return original weights

5. search for shortest tree (as in 1b)
6. return to 2. and repeat

> 30x faster than traditional search strategies

http://genealogyreligion.net/wp-content/uploads/2010/04/intelligencelandscape.jpg
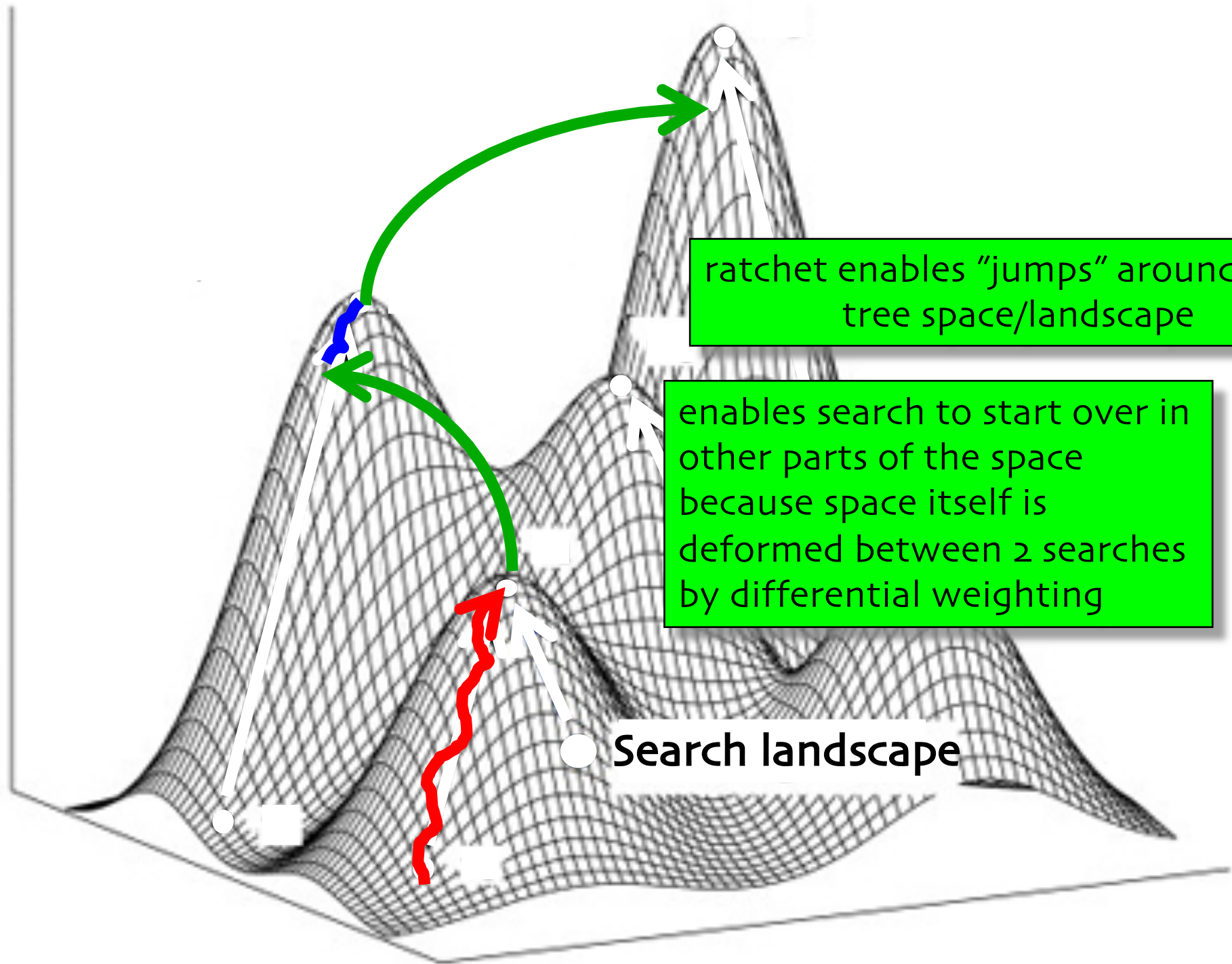
ratchet enables "jumps" around
tree space/landscape

enables search to start over in
other parts of the space
because space itself is
deformed between 2 searches
by differential weighting

Search landscape

SUMMARY

in order to find best hypotheses we should include into
our analyses ALL characters simultaneously

exhaustive & branch and bound searches can be used
only for analyses of SMALL matrices | only these two ENSURE
finding the optimal tree

heuristic search is based on rearrangement of branches
of tree(s) | branch SWAPPING |

searches should be planned carefully BEFORE starting
them in order to avoid unnecessary analyses of
large number of similar trees & use of CPU time