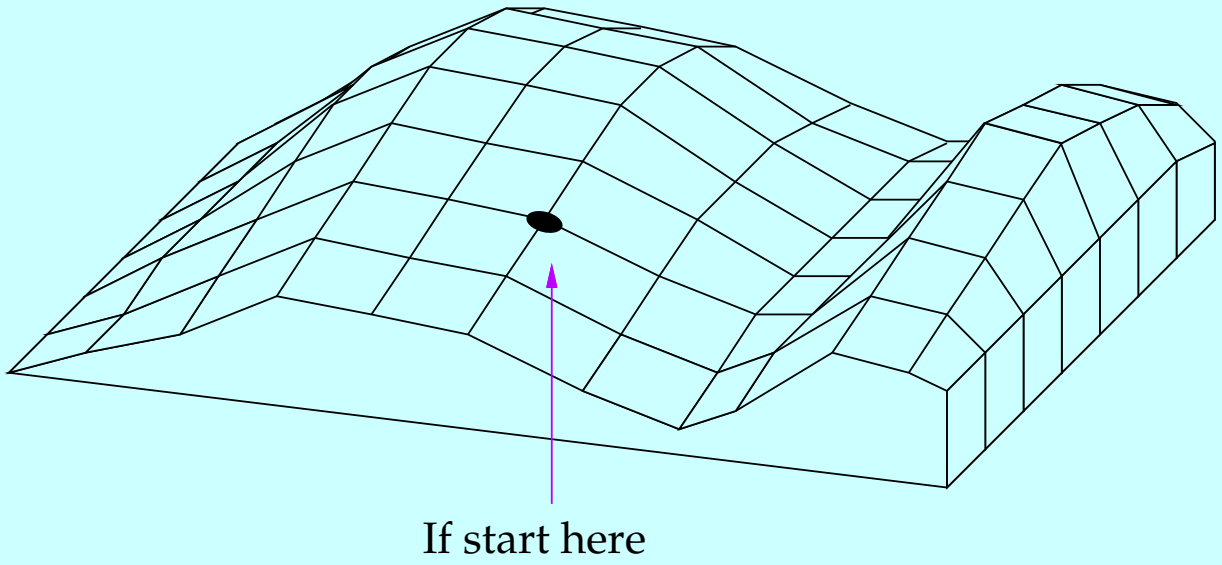
Week 2: Searching for trees, ancestral states

Genome 570

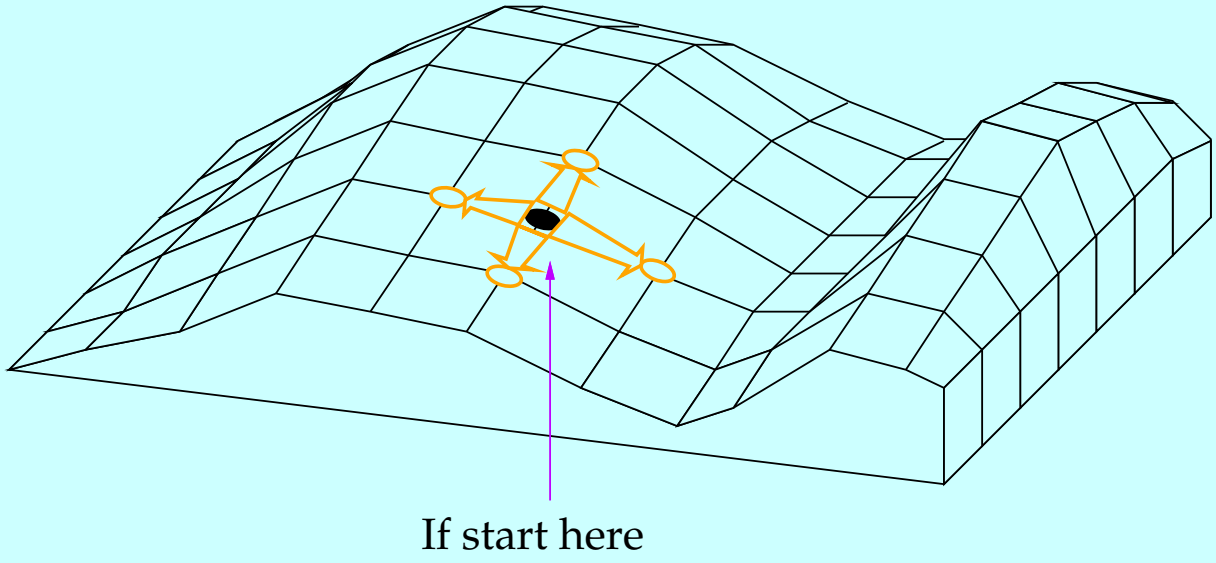
January, 2016

Greedy search for a maximum



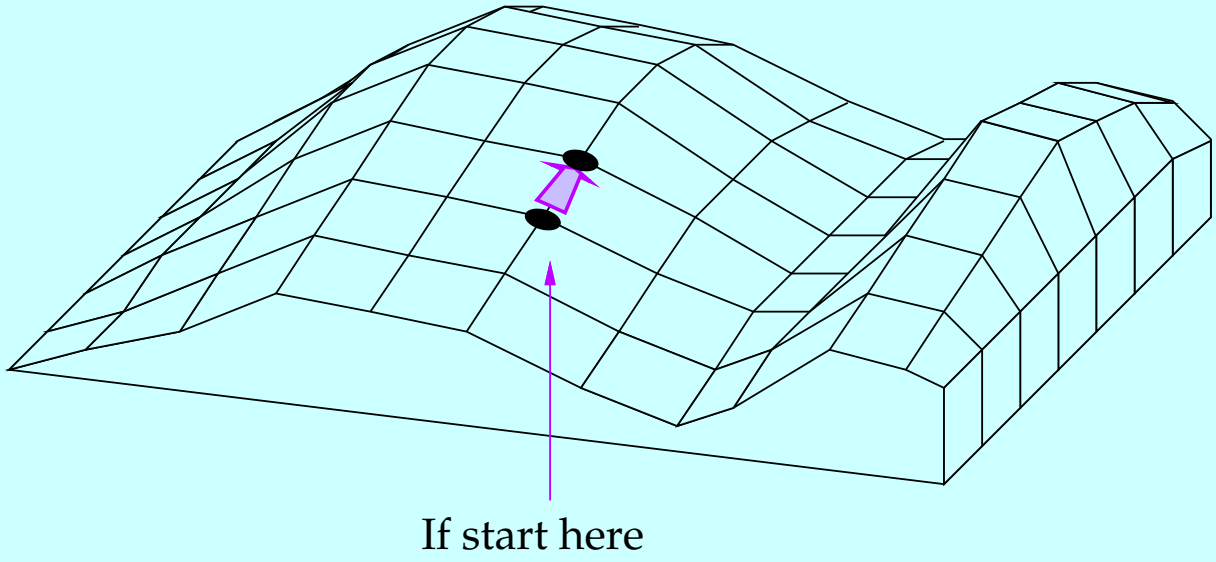
Parsimony methods search for a minimum. The surface is easier to see if we turn it upside down and search for a maximum. From an arbitrary starting point ...‘

Greedy search for a maximum



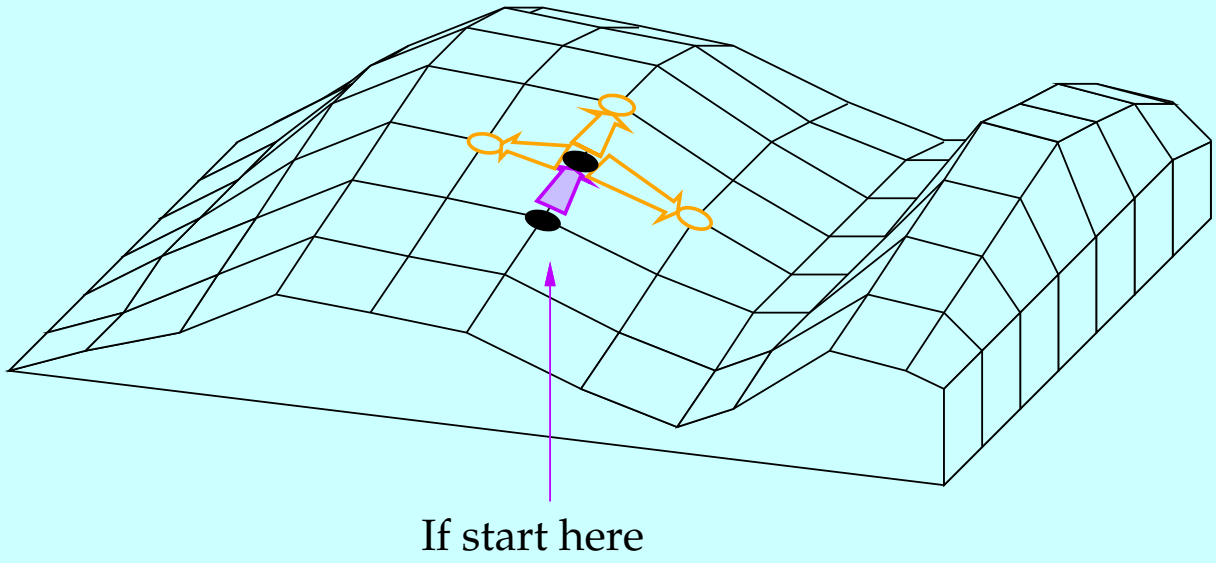
If we look at the neighboring points, and ...

Greedy search for a maximum



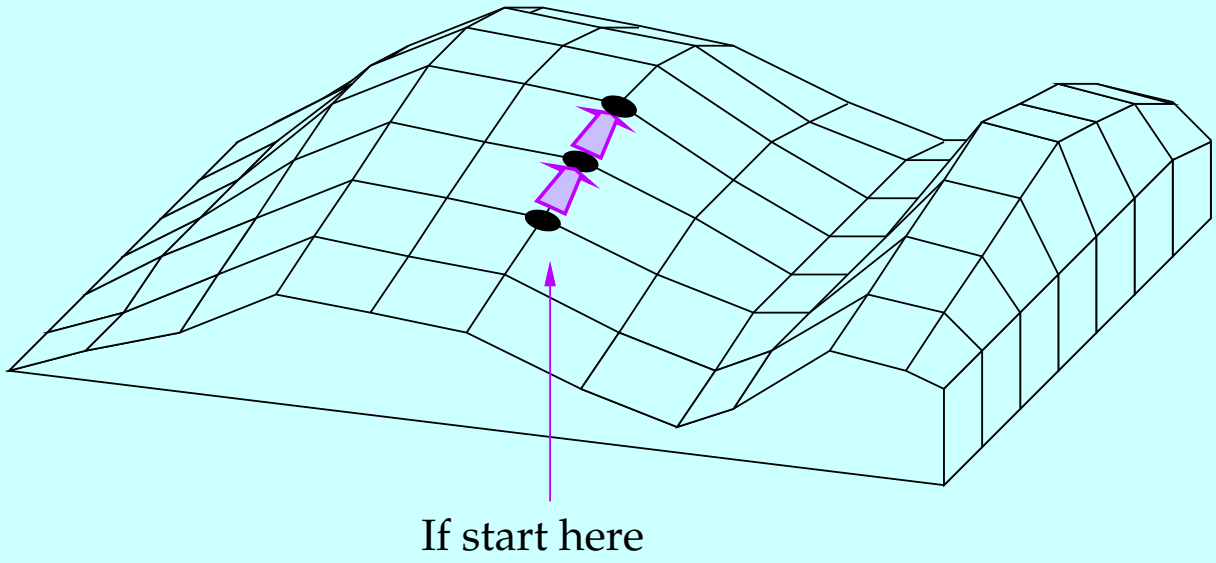
... then move to the highest one ...

Greedy search for a maximum



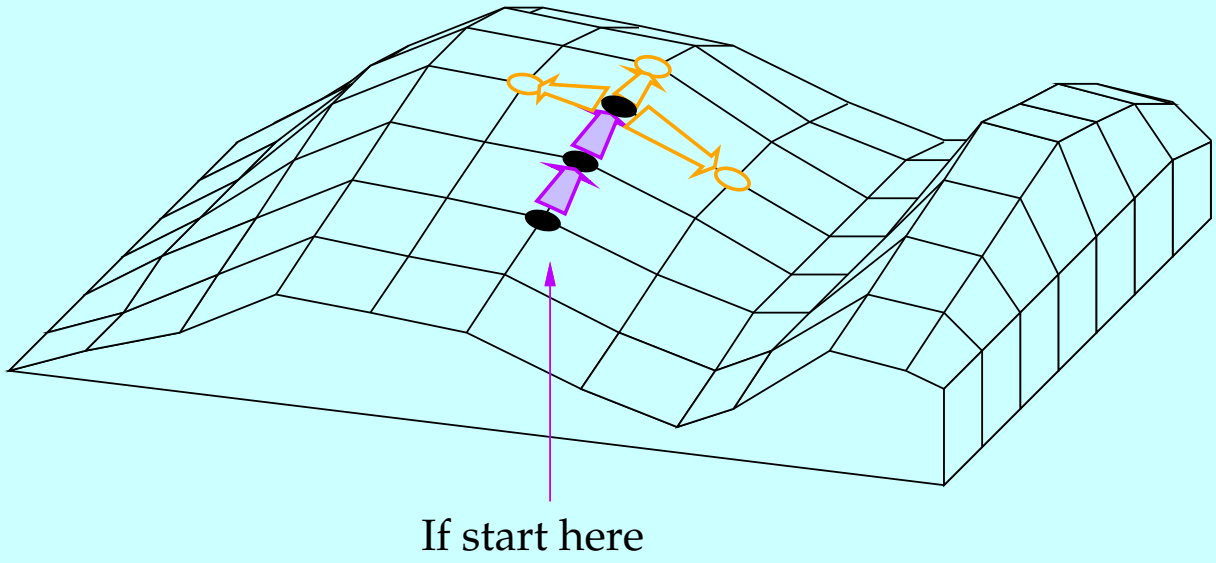
... looking at the neighboring points, and ...

Greedy search for a maximum



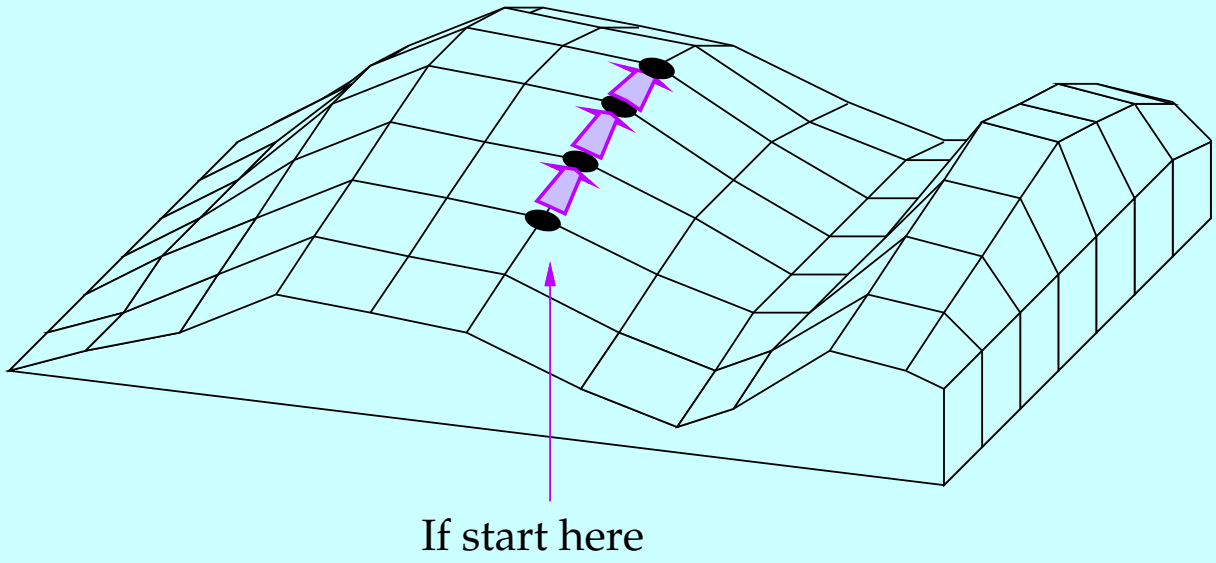
... then moving to the highest one,

Greedy search for a maximum



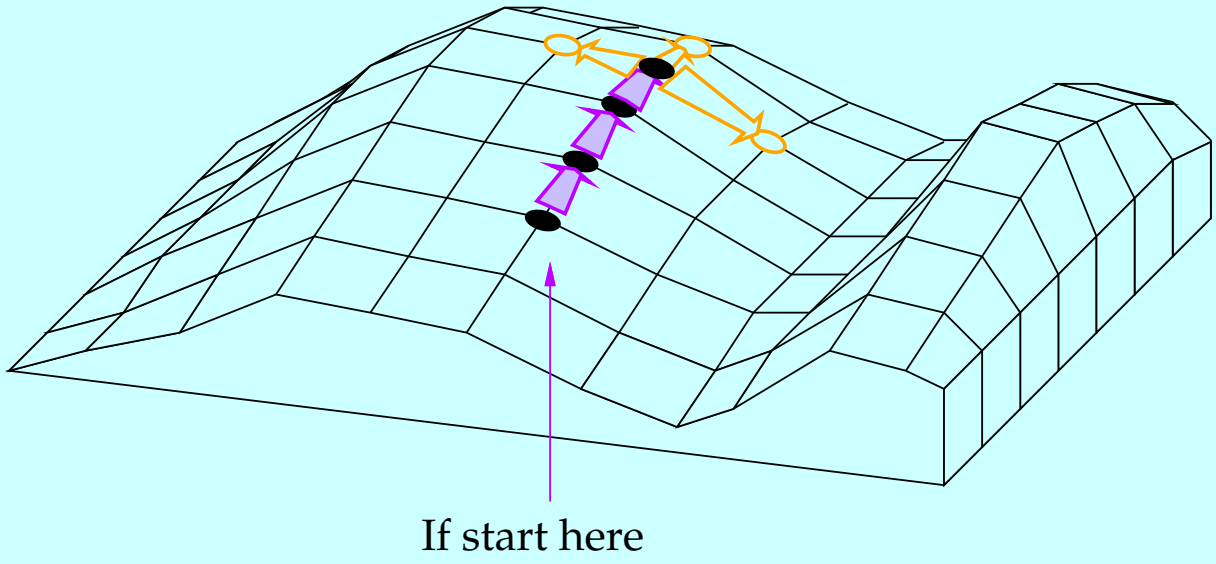
... looking at the neighboring points, and ...

Greedy search for a maximum



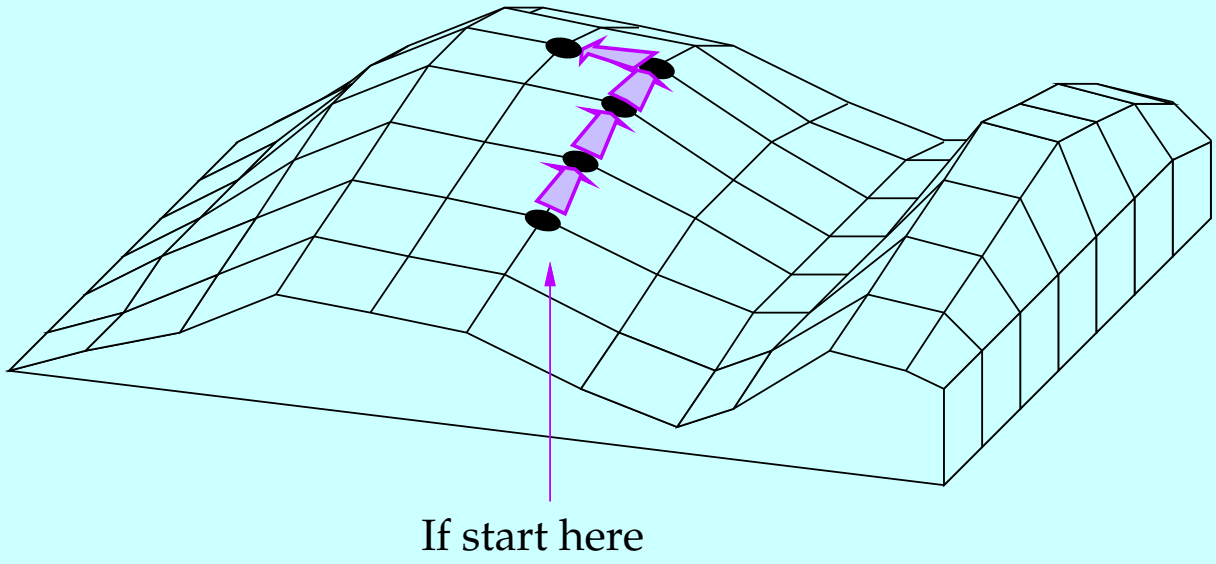
... then moving to the highest one,

Greedy search for a maximum



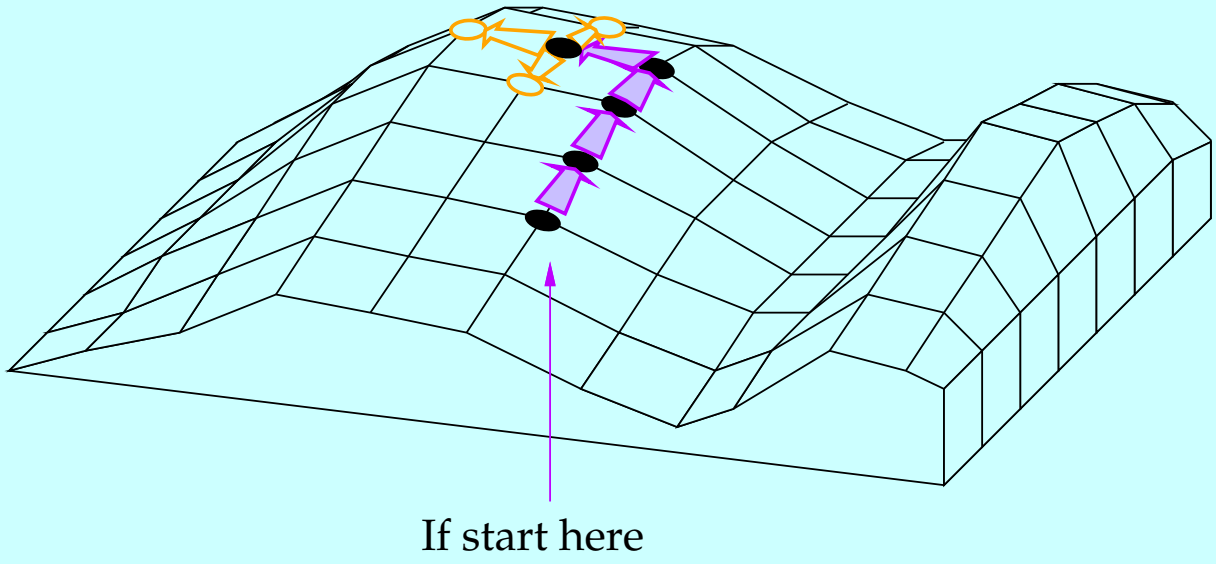
... looking at the neighboring points, and ...

Greedy search for a maximum



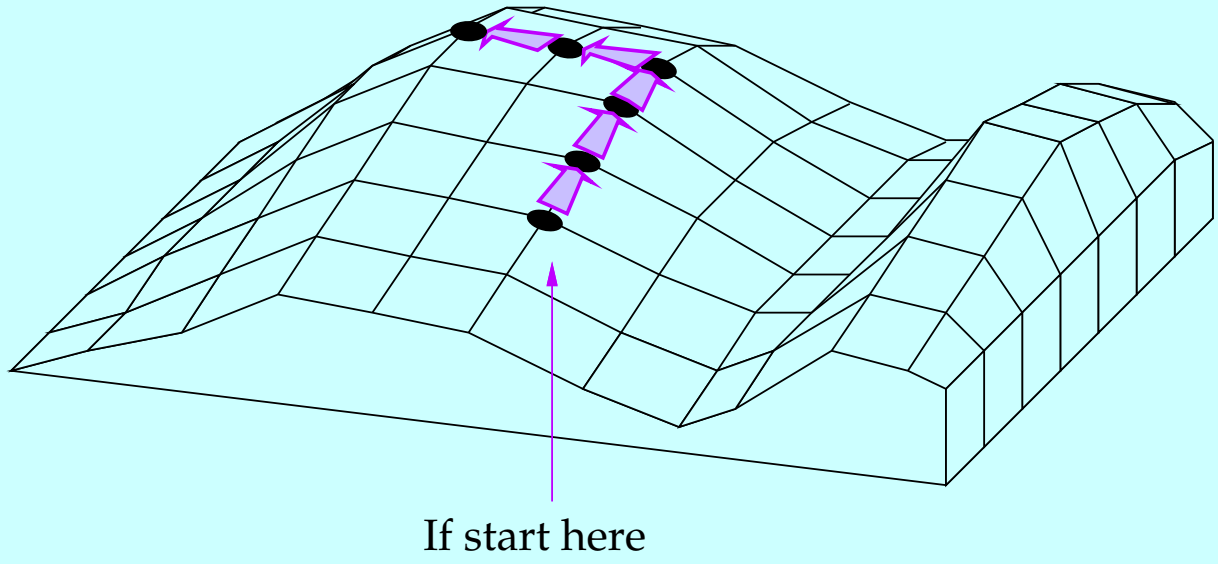
... then moving to the highest one,

Greedy search for a maximum



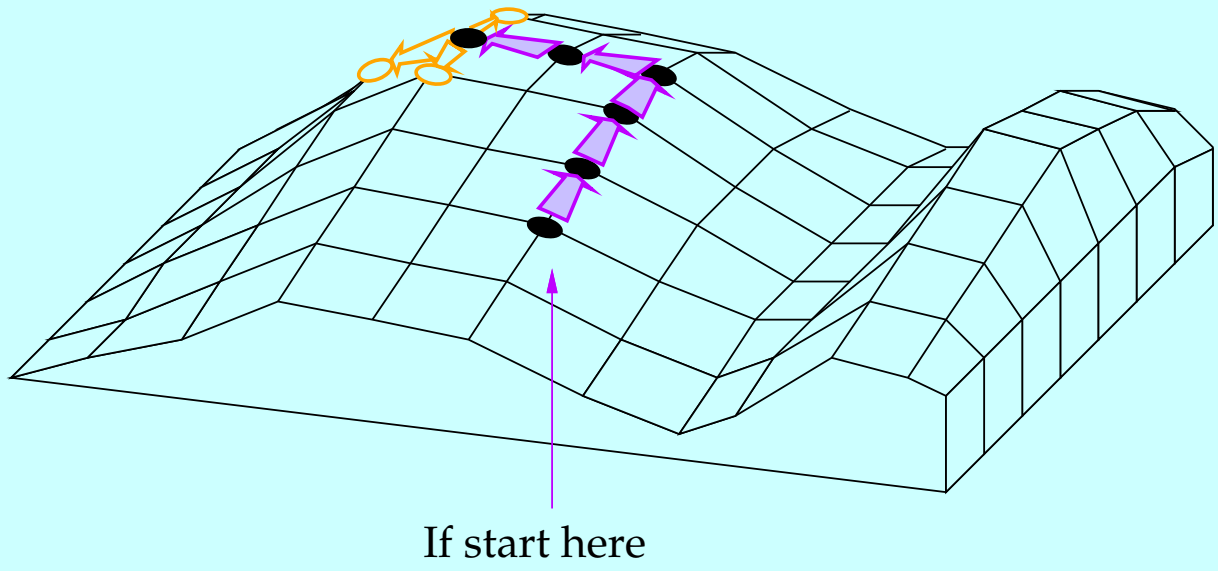
... looking at the neighboring points, and ...

Greedy search for a maximum



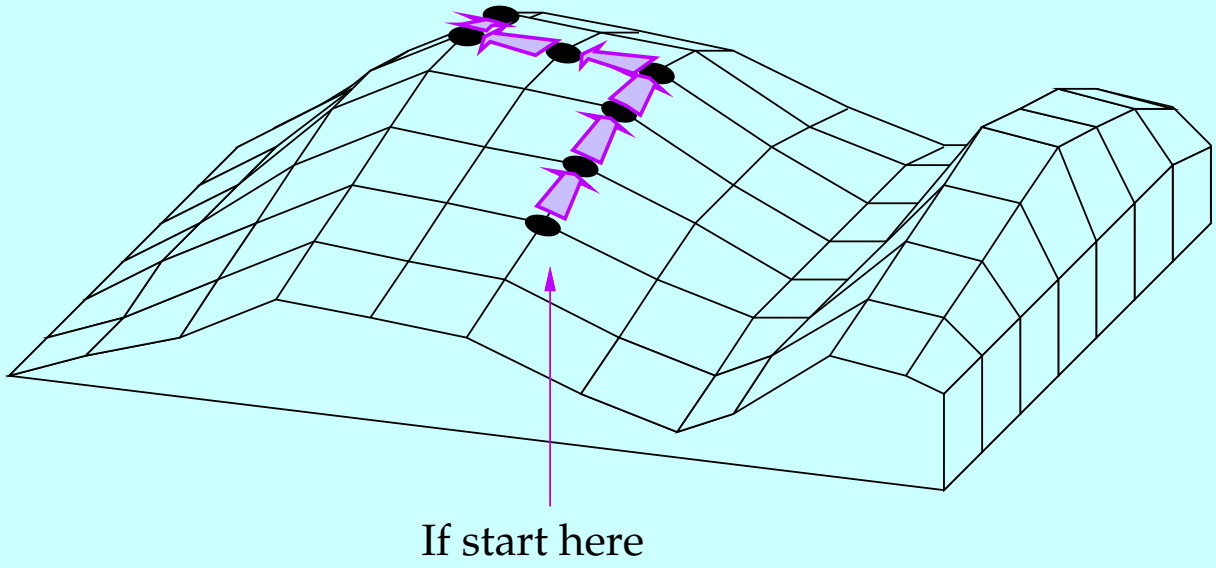
... then moving to the highest one,

Greedy search for a maximum



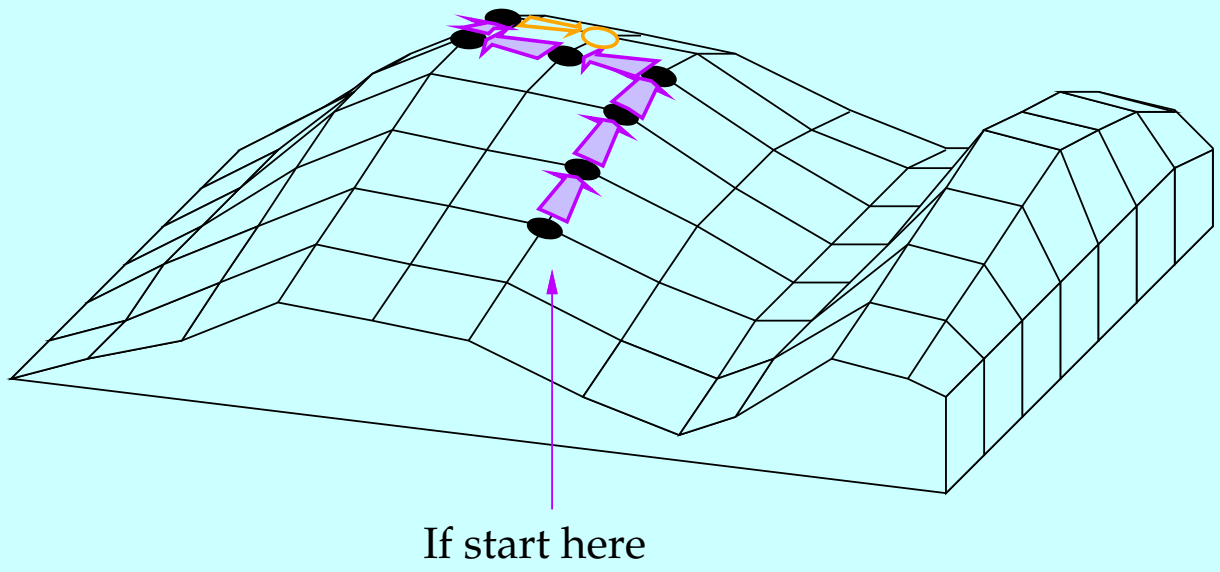
... looking at the neighboring points, and ...

Greedy search for a maximum



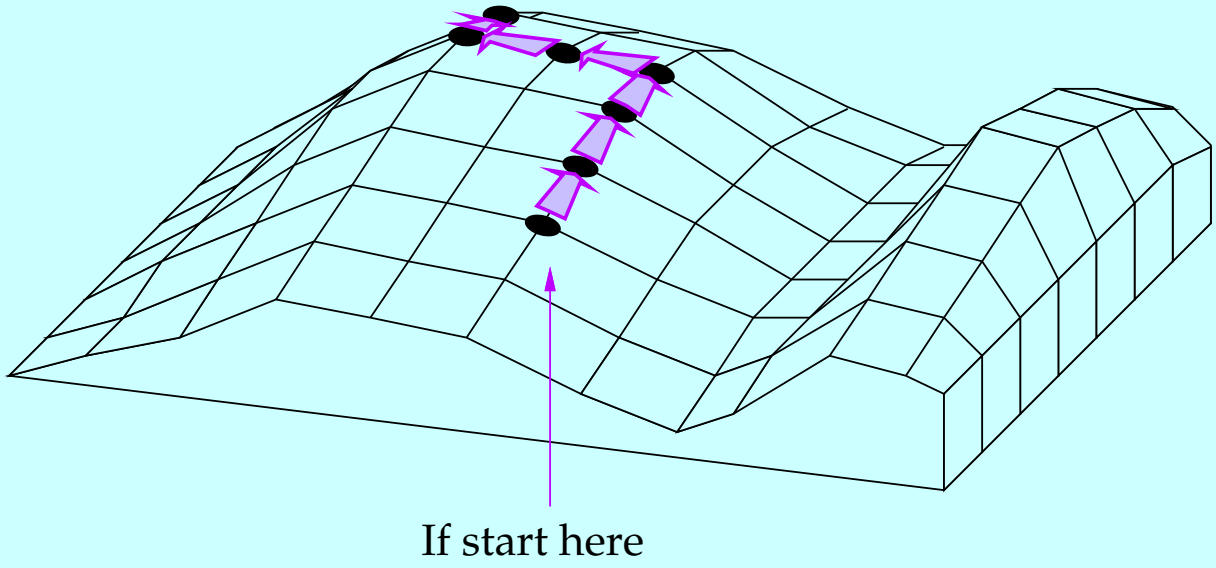
... then moving to the highest one,

Greedy search for a maximum



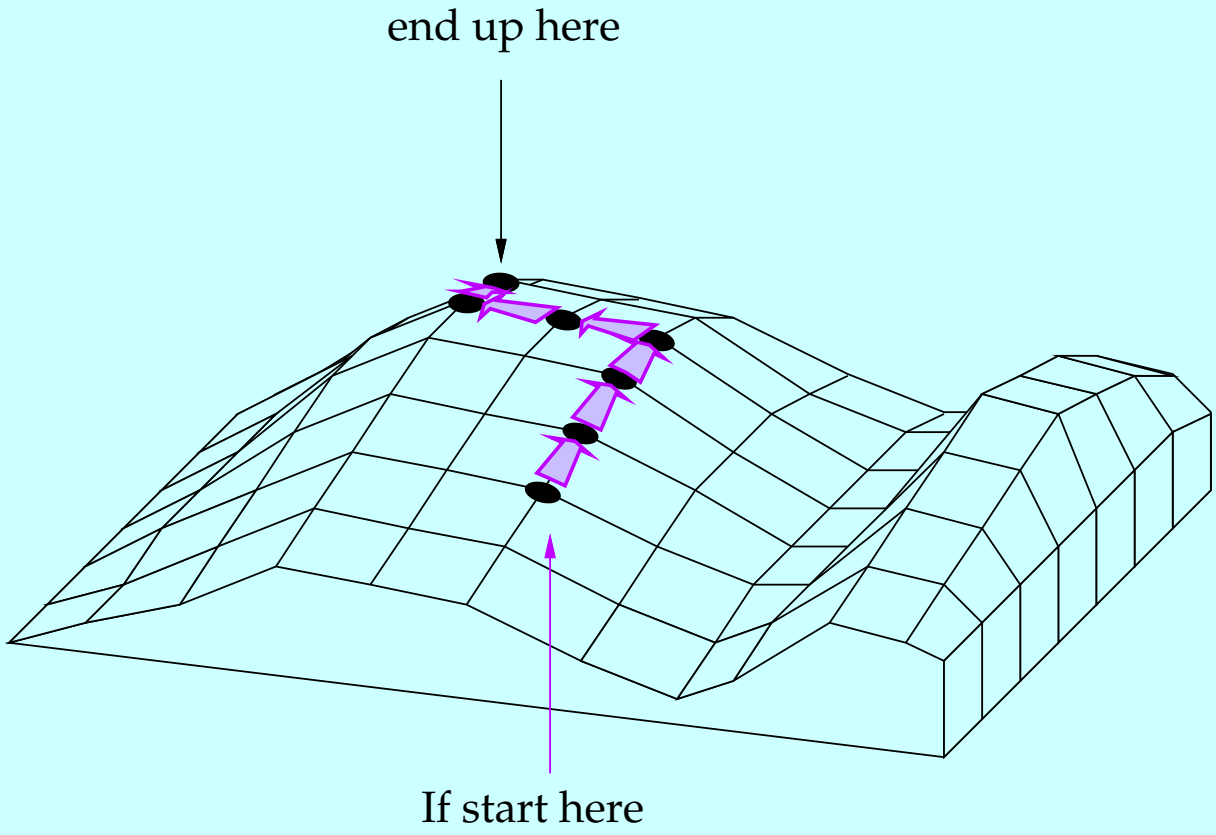
... until, looking at the neighboring points, we find none that are better ...

Greedy search for a maximum



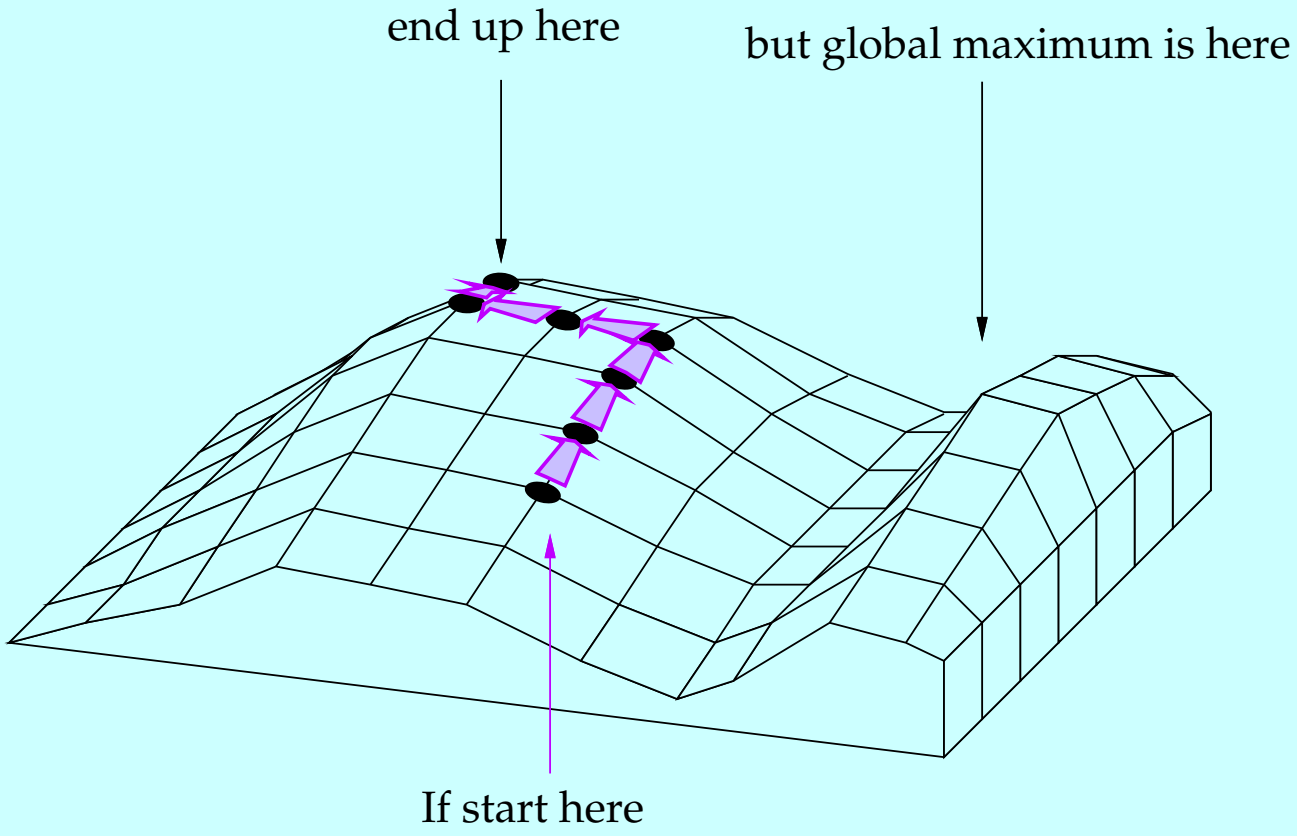
... so we stop there, ...

Greedy search for a maximum



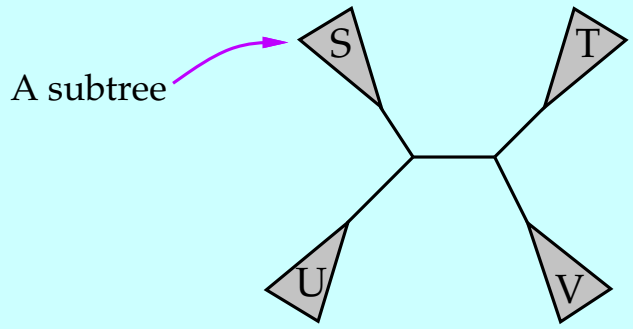
... we will find better points ...

Greedy search for a maximum

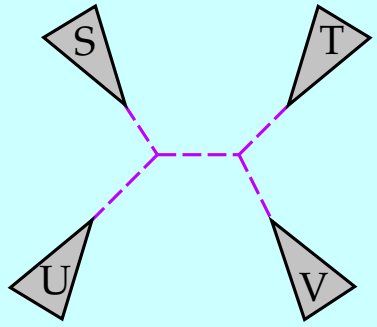


... but not necessarily the overall best point.

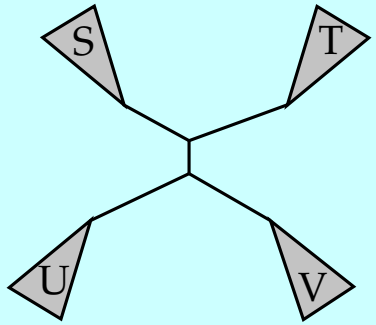
Nearest-neighbor interchange (NNI) rearrangements

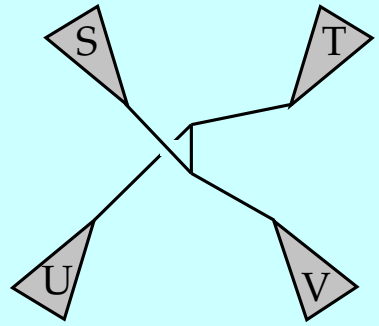


is rearranged by dissolving the connections to an interior branch

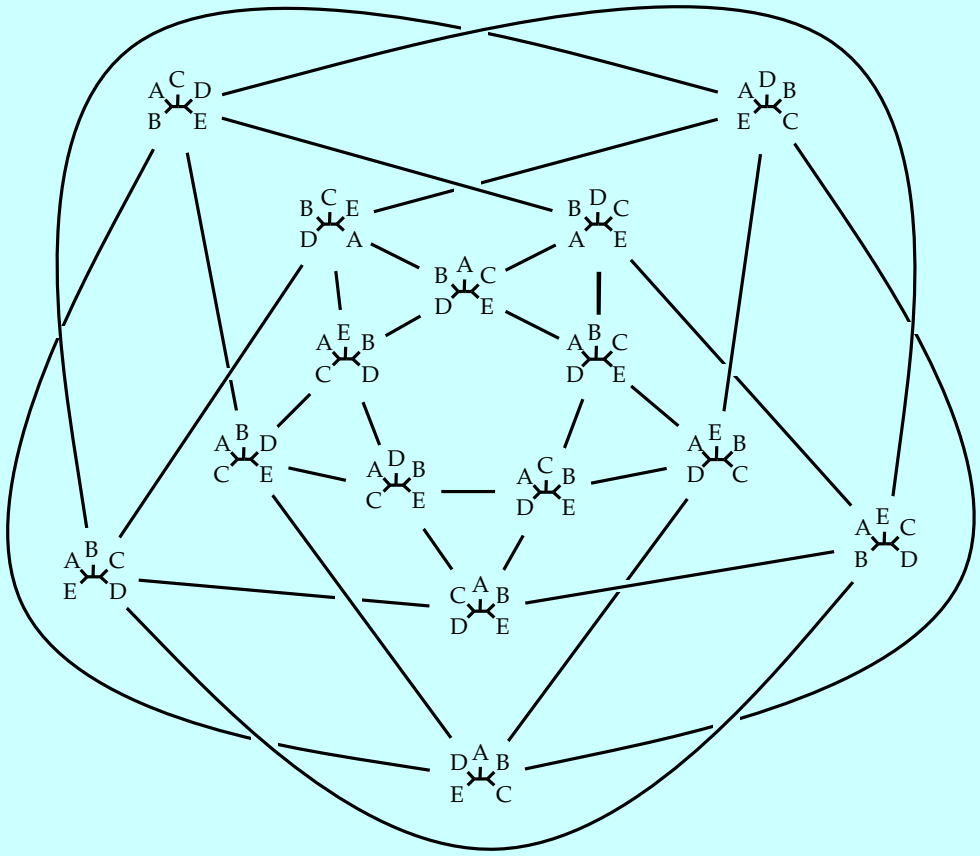


and reforming them in one of the two possible alternative ways:





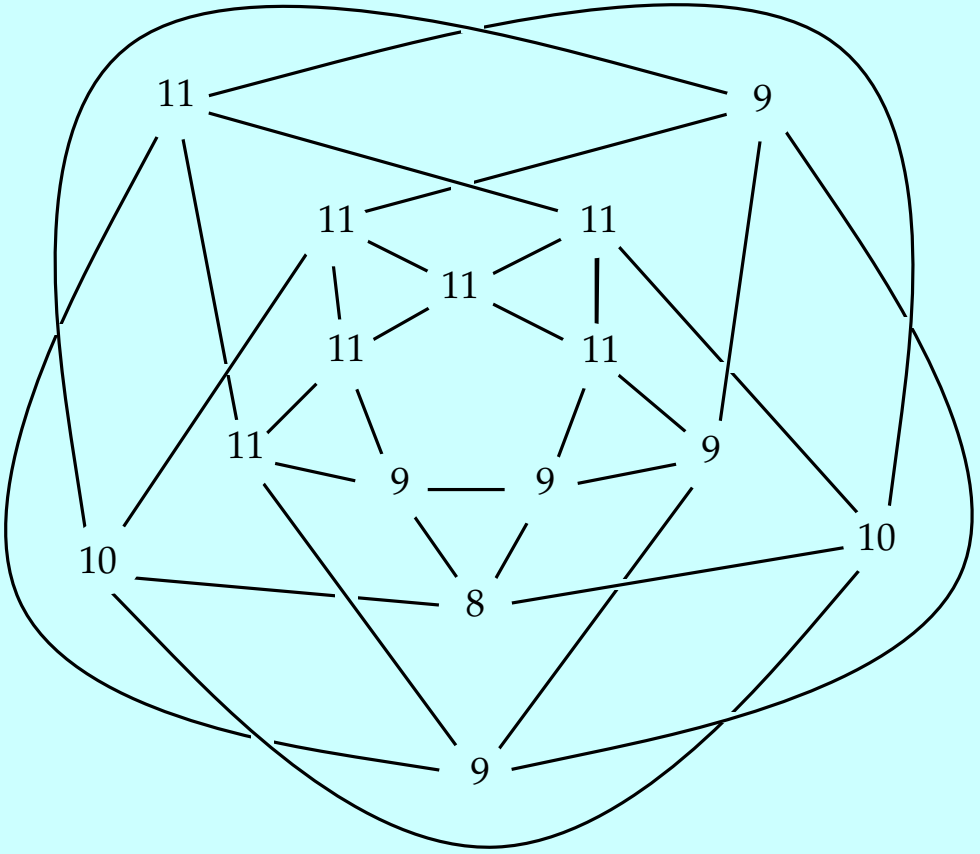
The Schoenberg graph



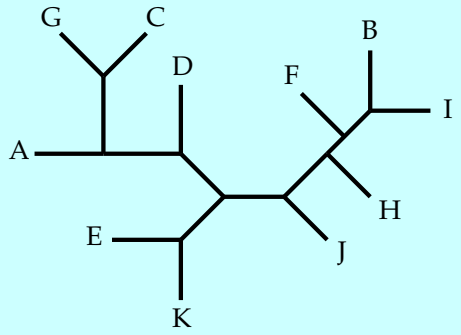
All 15 unrooted 5-species trees, connected if a Nearest-Neighbor Interchange can take you from one to the other.

(This arrangement of the graph is due to Ben Schoenberg).

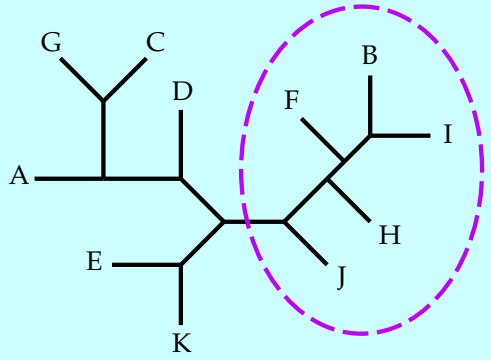
With numbers of steps of trees



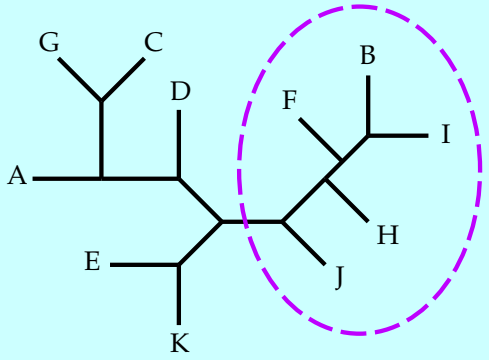
Subtree pruning and regrafting (SPR)



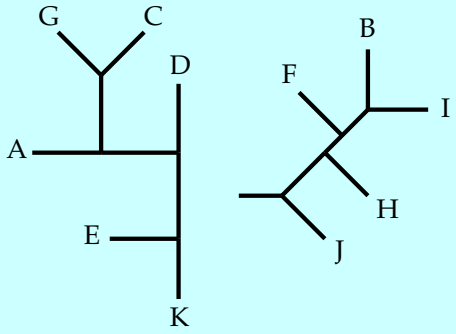
Subtree pruning and regrafting



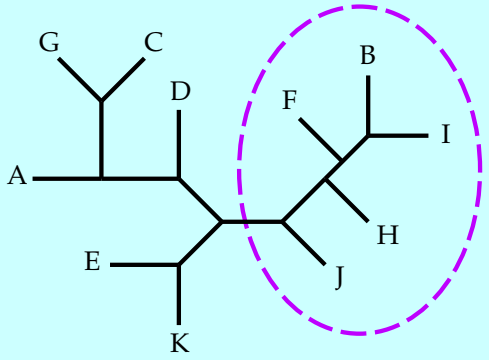
Subtree pruning and regrafting



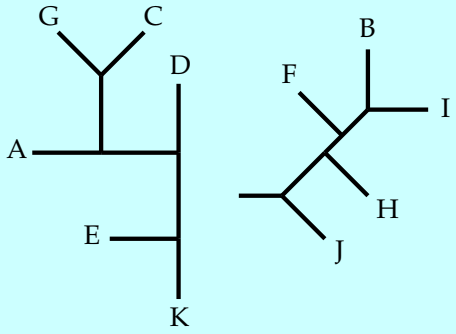
Break a branch, remove a subtree



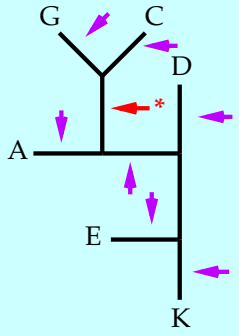
Subtree pruning and regrafting



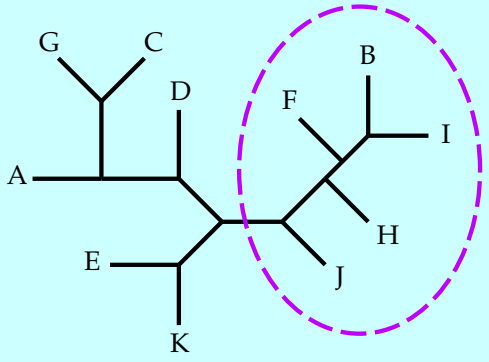
Break a branch, remove a subtree



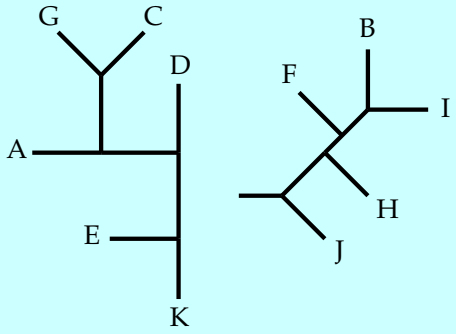
Add it in, attaching it to one (*)
of the other branches



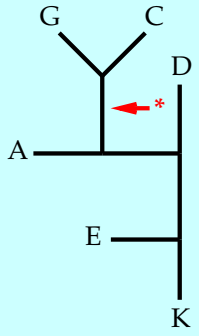
Subtree pruning and regrafting



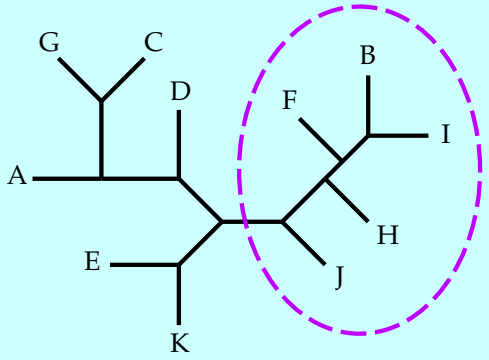
Break a branch, remove a subtree



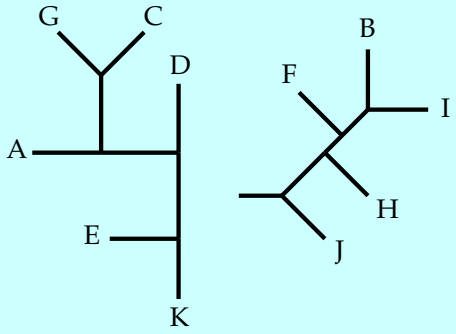
Add it in, attaching it to one (*)
of the other branches



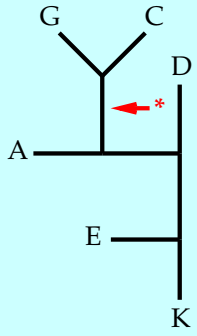
Subtree pruning and regrafting



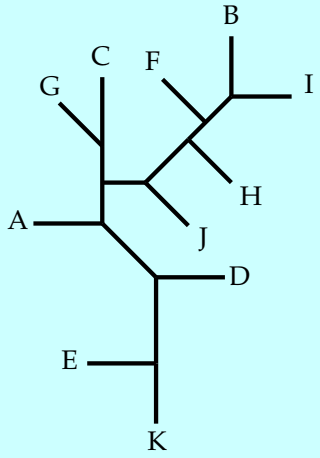
Break a branch, remove a subtree



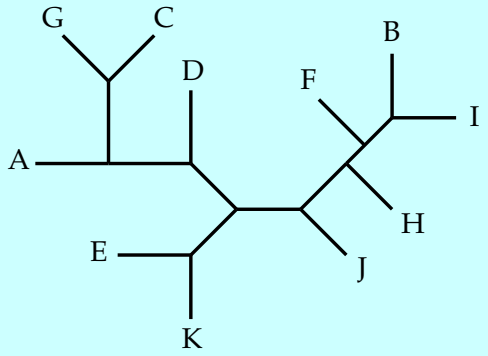
Add it in, attaching it to one (*)
of the other branches



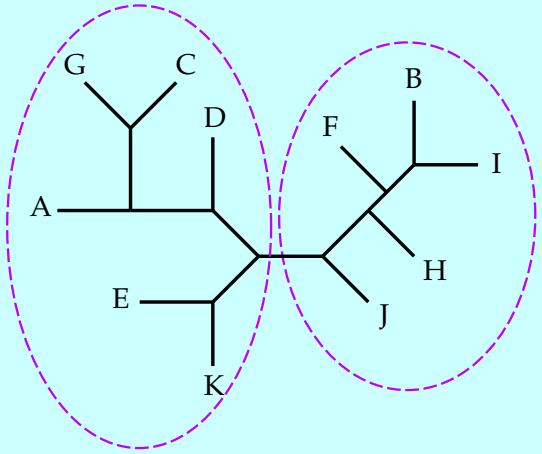
Here is the result:



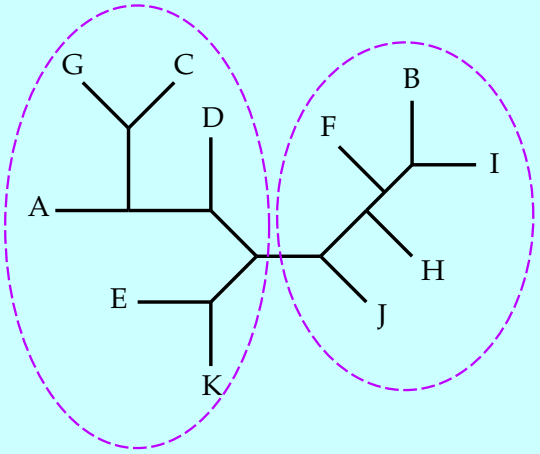
Tree bisection and reconnection (TBR)



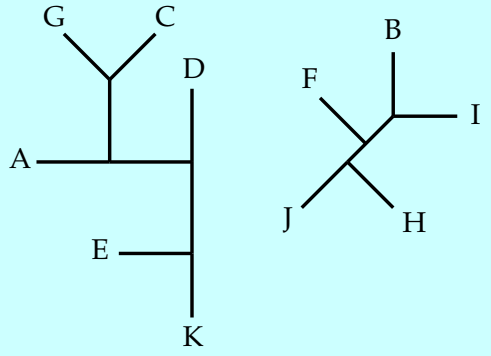
Tree bisection and reconnection



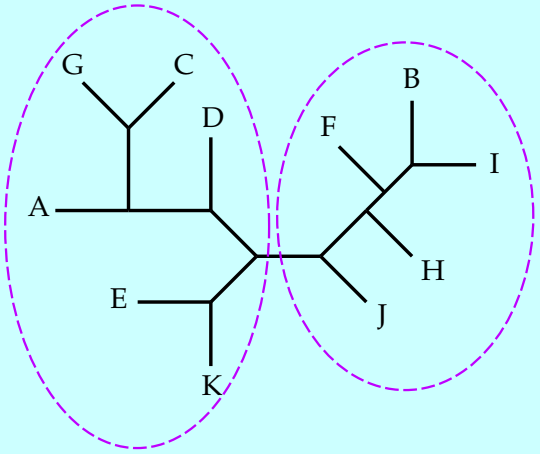
Tree bisection and reconnection



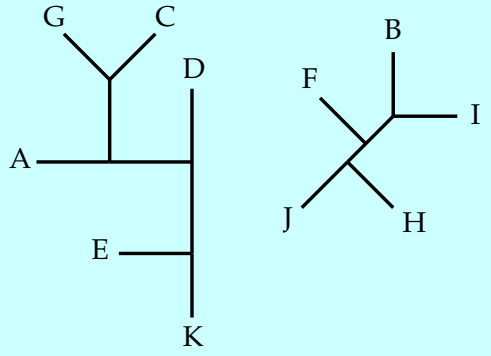
Break a branch, separate the subtrees



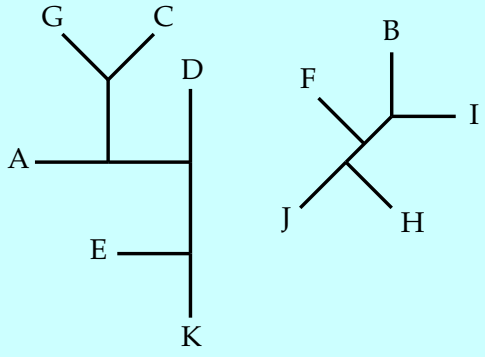
Tree bisection and reconnection



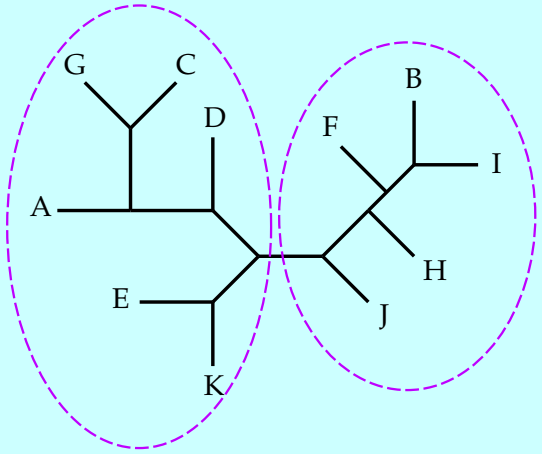
Break a branch, separate the subtrees



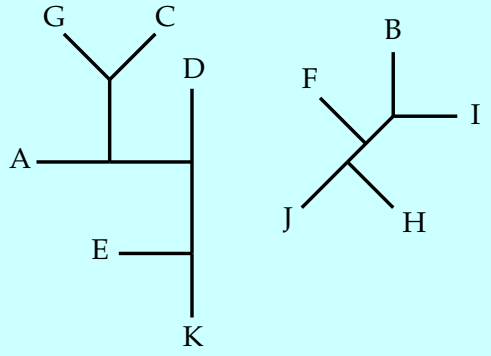
Connect a branch of one
to a branch of the other



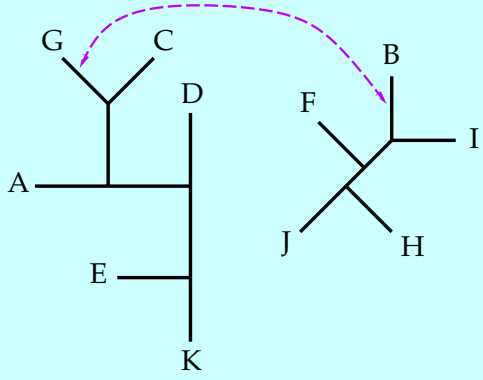
Tree bisection and reconnection



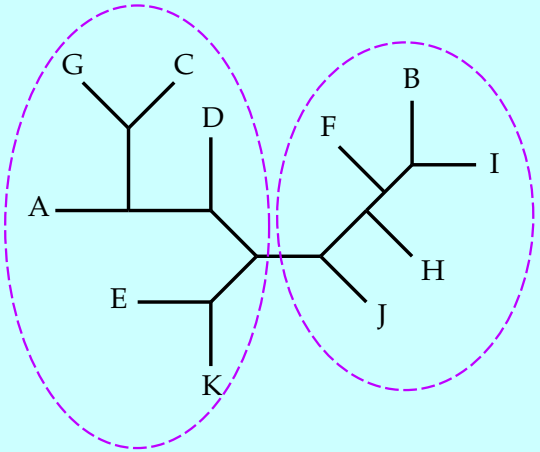
Break a branch, separate the subtrees



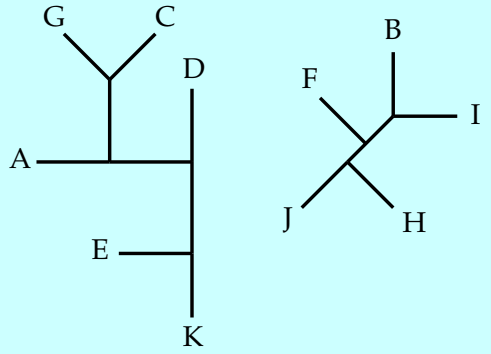
Connect a branch of one
to a branch of the other



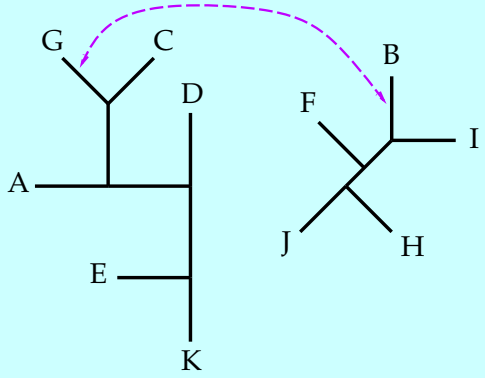
Tree bisection and reconnection



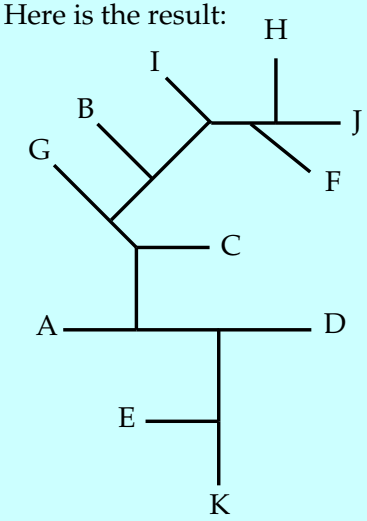
Break a branch, separate the subtrees



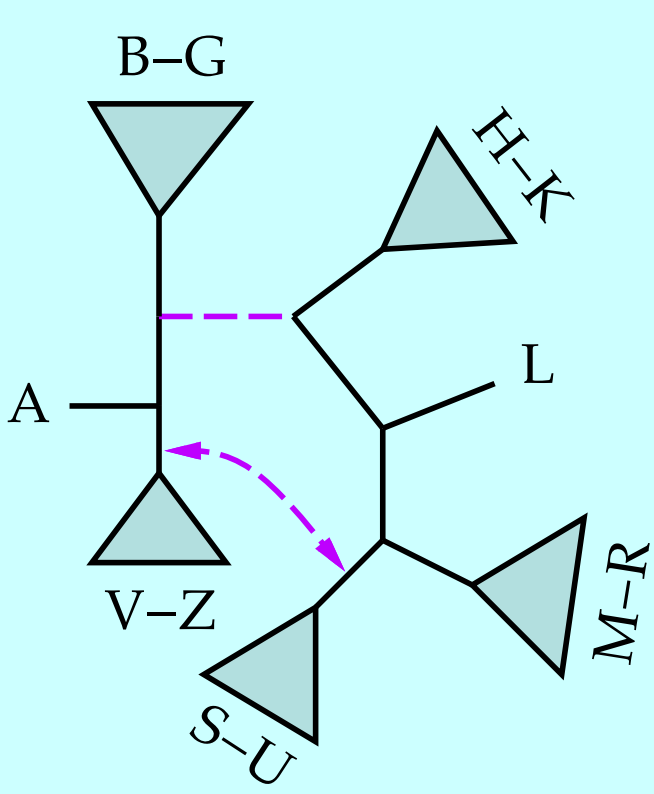
Connect a branch of one
to a branch of the other

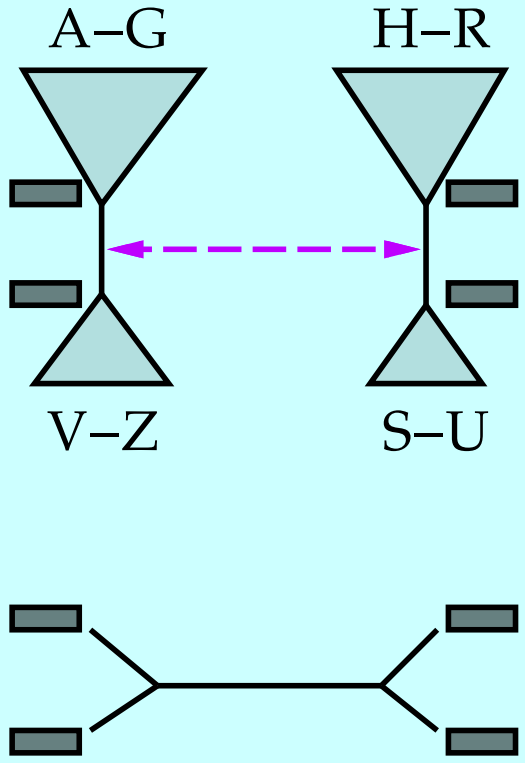


Here is the result:



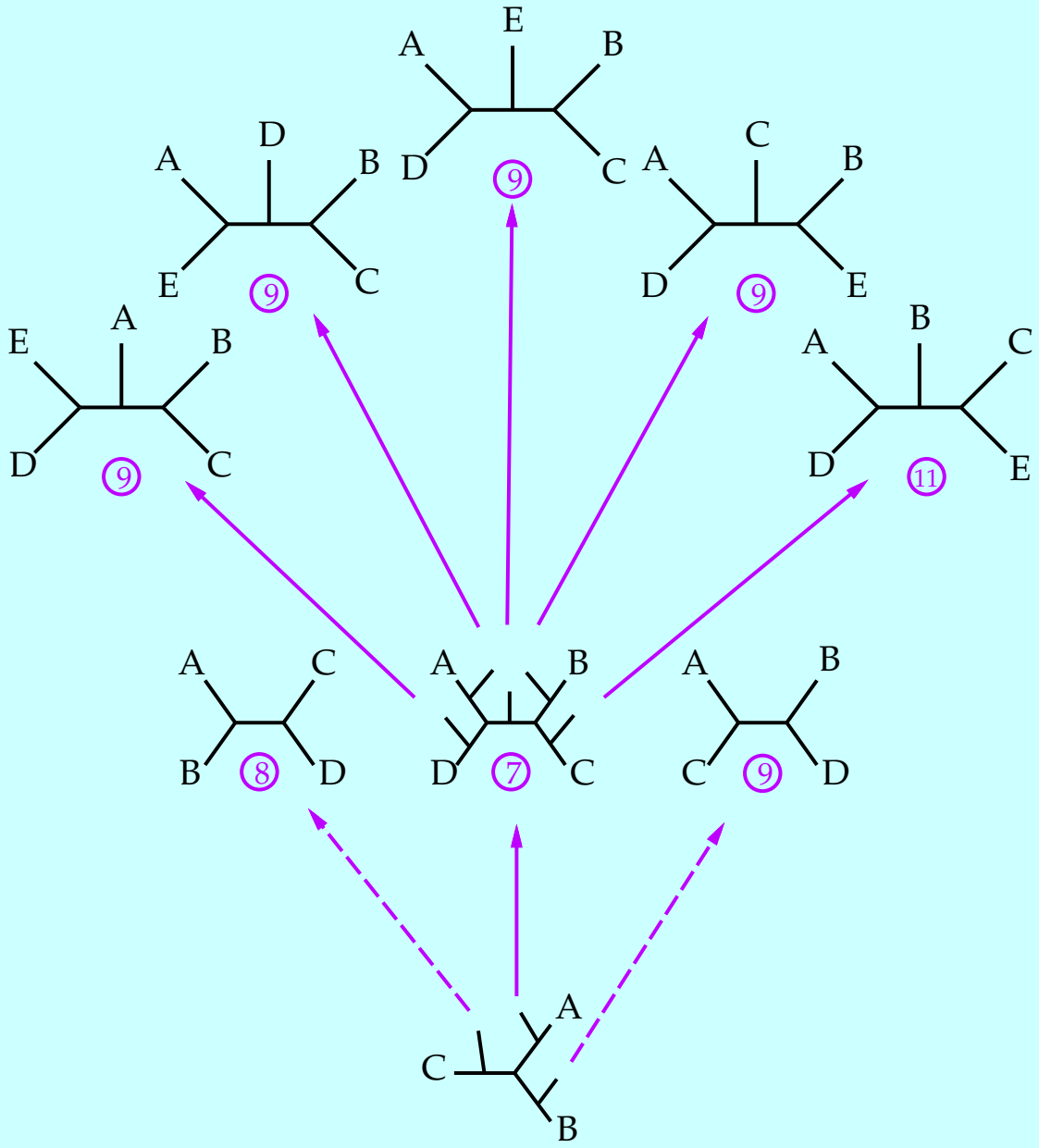
Goloboff's economy in rearranging



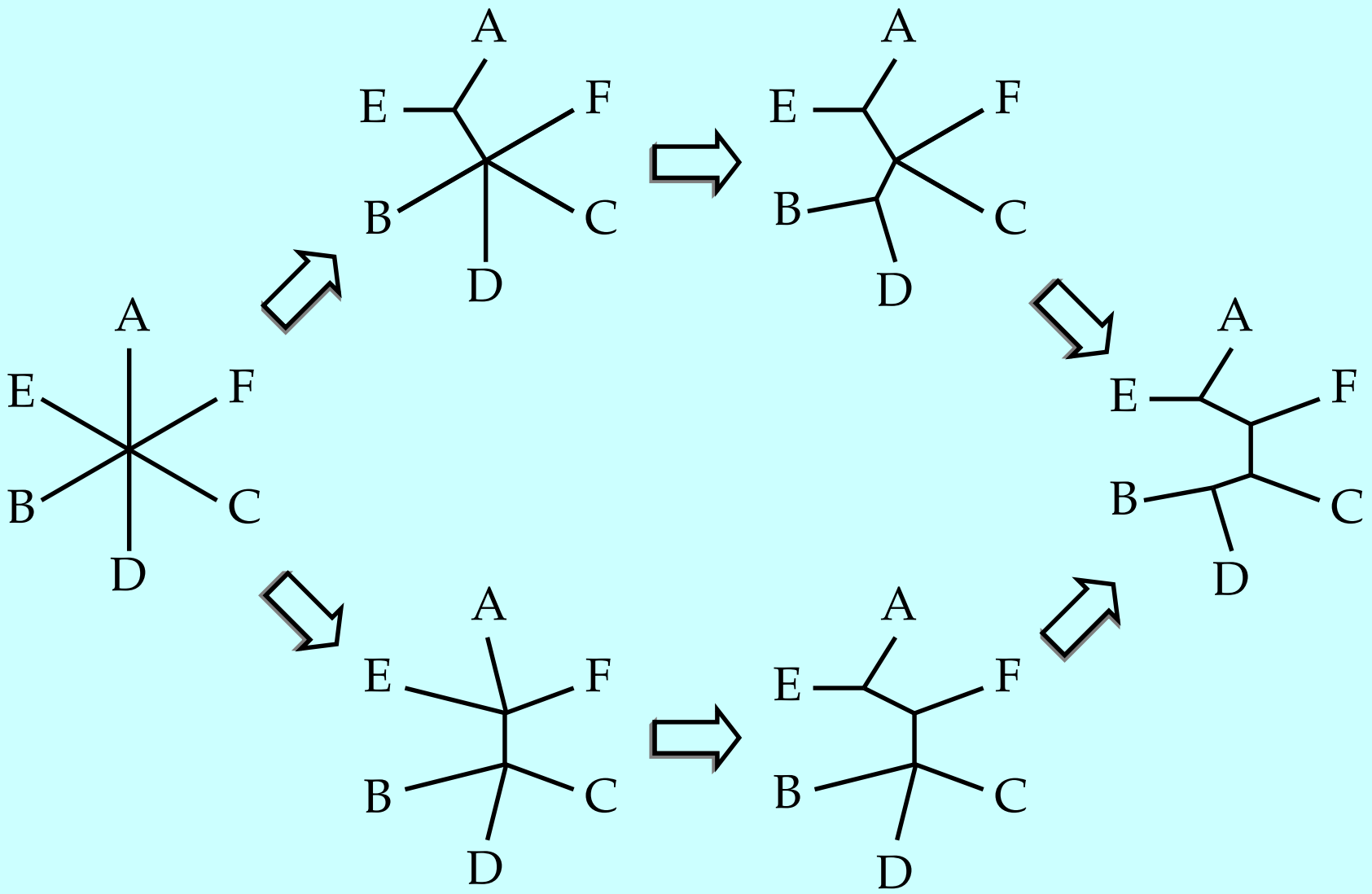


If we compute the relevant interior arrays of steps beyond that point in the tree, we can very quickly evaluate different reconnections of the two parts of the tree.

Sequential addition



Star decomposition



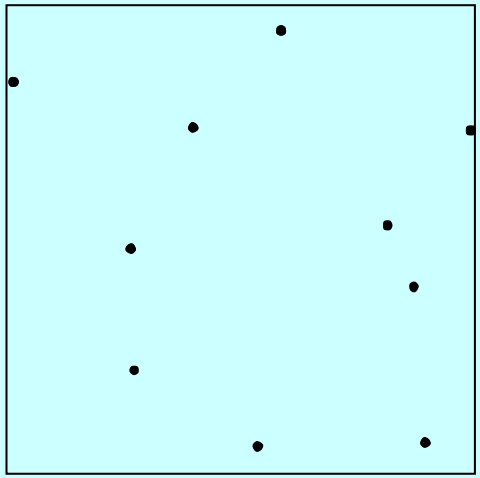
Some cleverer rearrangement methods

- Kevin Nixon's "parsimony ratchet" search (*Cladistics*, 1999). This samples subsets of characters, uses just those characters in heuristic search to find a tree, then evaluates the resulting tree on the full data set. If it is better than the previous best tree, it becomes the starting point for more heuristic search, and more rounds of the ratchet. This works for other criteria other than parsimony too.

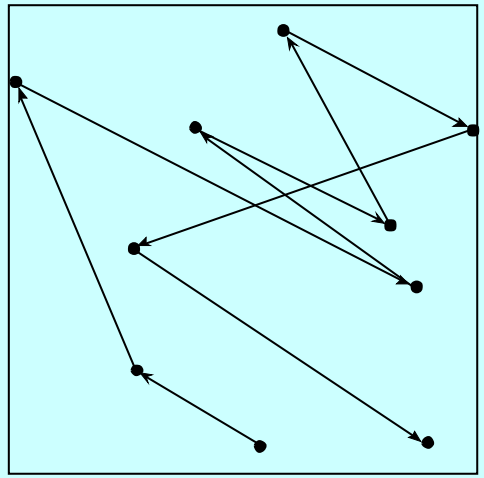
- Genetic algorithms. Several researchers have made genotypes that each describe a tree, and searched by computing a fitness from the parsimony score of the tree, and evolving a population of trees based on mutation, mating, selection, and recombination among these tree representation "genotypes".

An example of looking for the Shortest Hamiltonian path

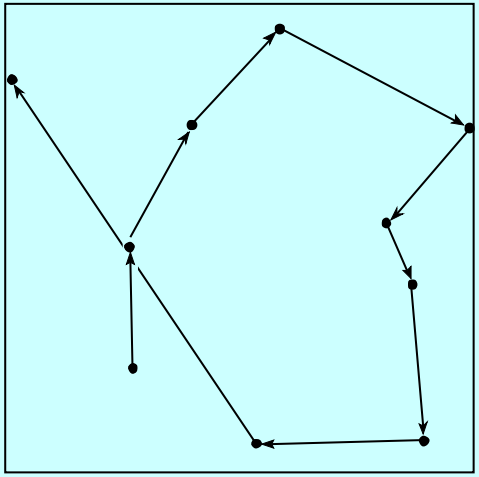
10 random points



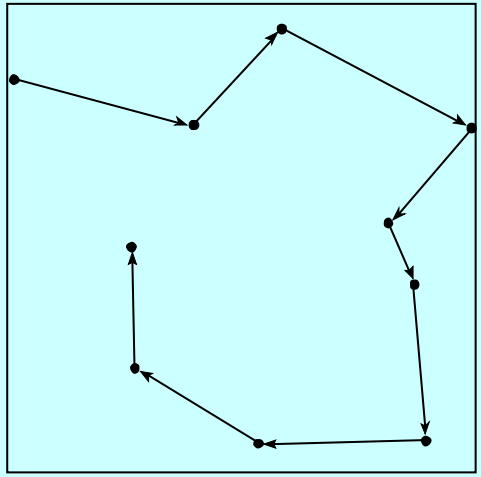
A random path



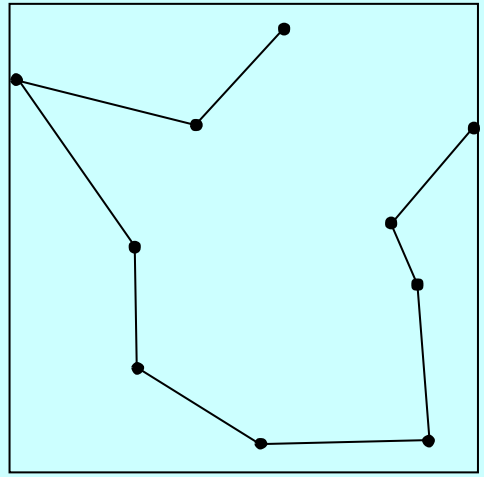
One greedy path



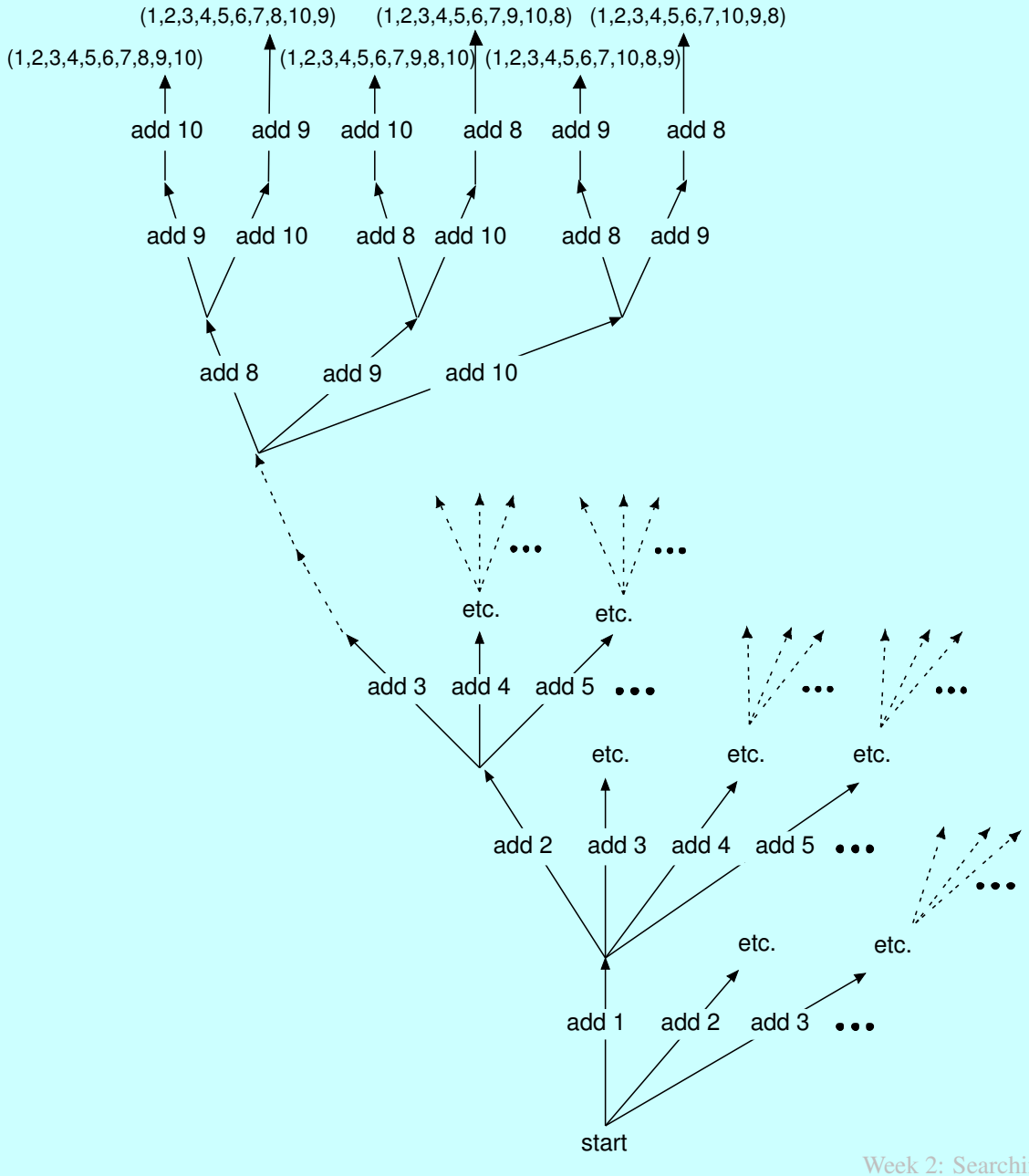
Best greedy path



Best path



A search tree of paths

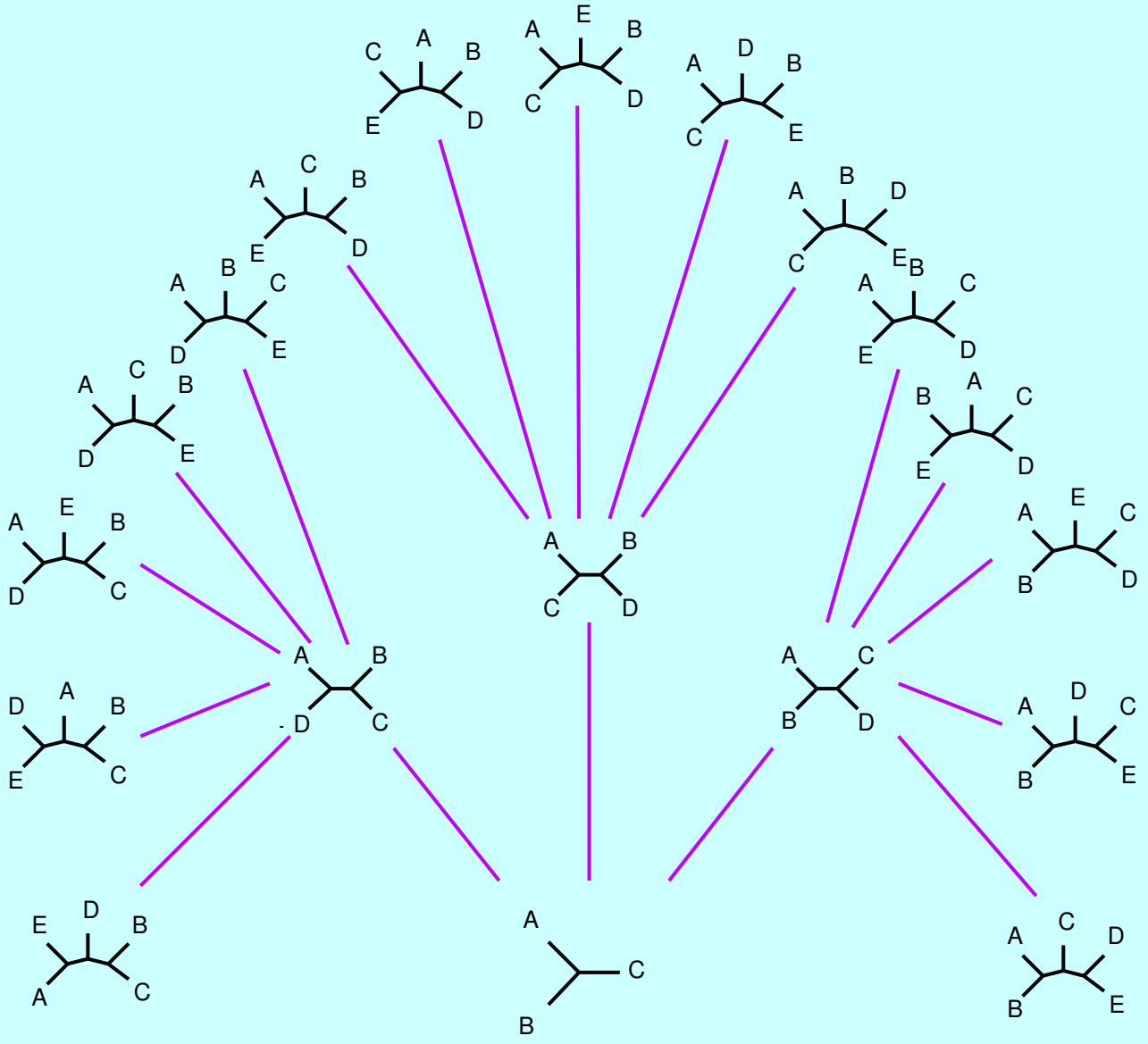


Time-savings of branch and bound

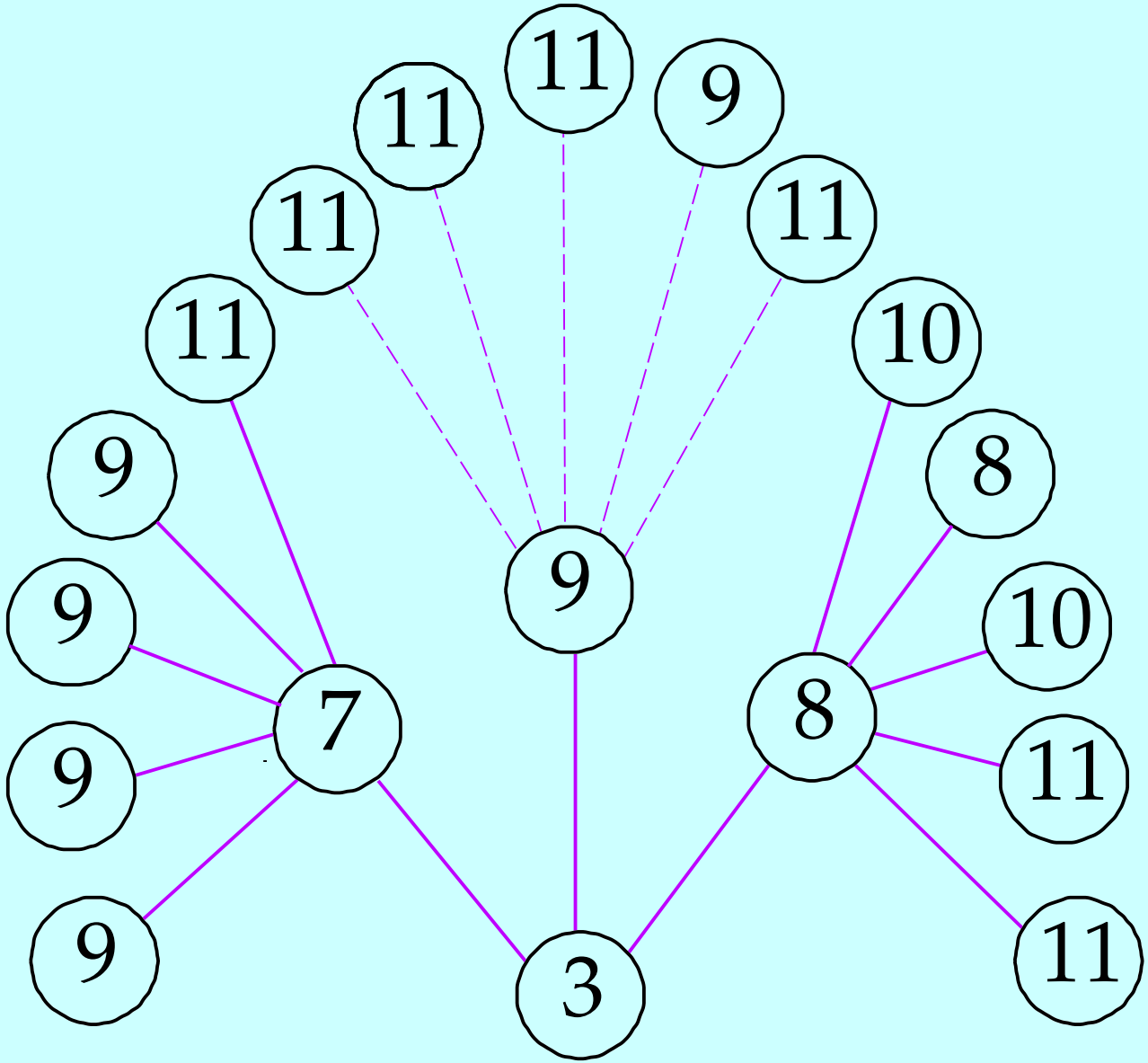
Results for this case:

Algorithm	length	time
Greedy search from all points	2.802660	(too fast to measure)
Exhaustive enumeration	2.781230	10.85 sec
Branch and bound	2.781230	0.46 sec

Branch and bound on trees



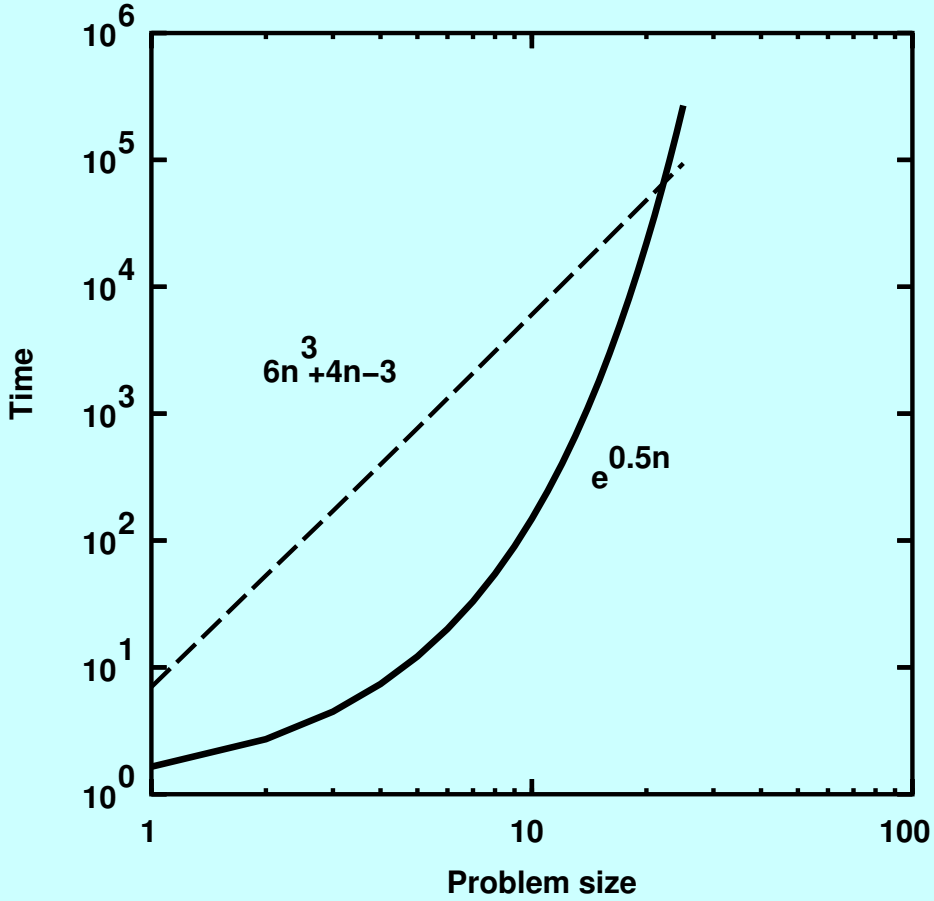
Branch and bound using numbers of steps



Calculating a lower bound on tree score

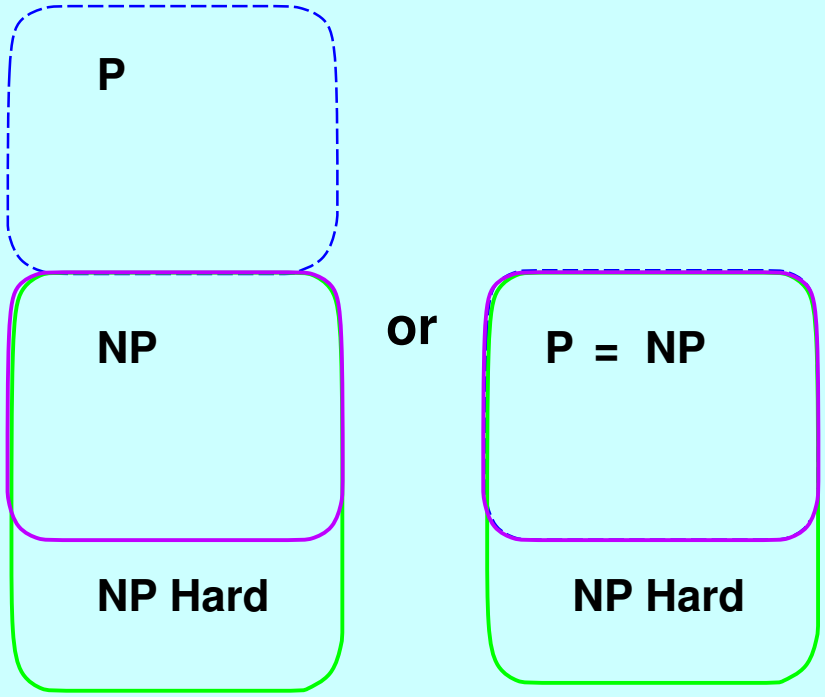
- The score of the partial tree is a lower bound (since adding more species cannot decrease the number of steps)
- Also can add the number of characters that do not show variation on the species added so far, but will once added (actually, the number of new states that will appear once all species are added – if A and G are there already, will C also appear?)
- Can also take all disjoint pairs of characters that will become incompatible once added, but aren't incompatible now (this is due to Dave Swofford. Each brings in one more step.)

Polynomial time and exponential time



How does the time taken by an algorithm depend on the size of the problem? If it is a polynomial (even one with big coefficients), with a big enough case it is faster than one that depends on the size exponentially.

NP completeness and NP hardness

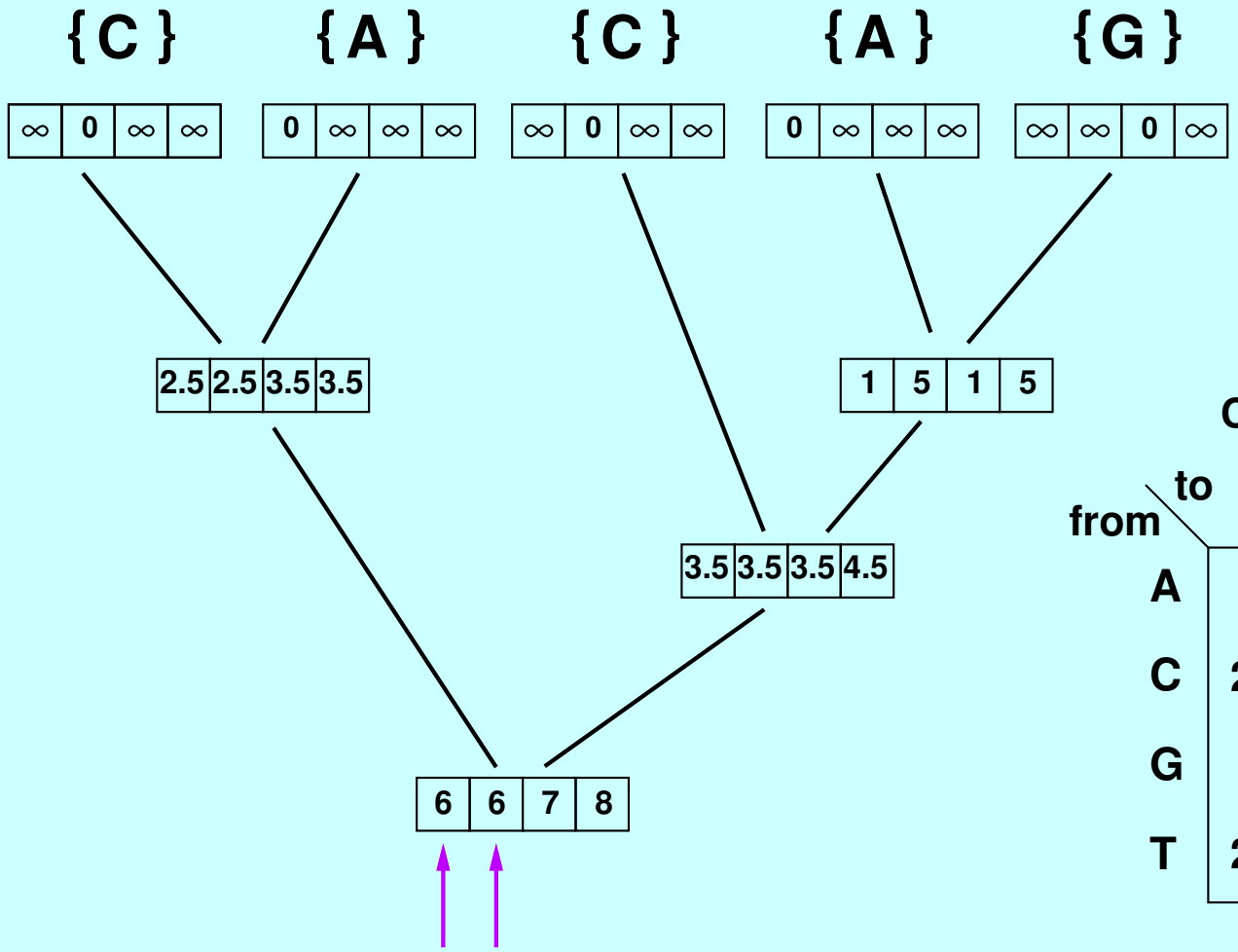


P = problems that can be solved by a polynomial time algorithm

NP complete = problems for which a proposed solution can be checked in polynomial time but for which it can be proven that if one of them is in P, all of them are.

NP hard = problems for which a solution can be checked in polynomial time, but might be not solvable in polynomial time

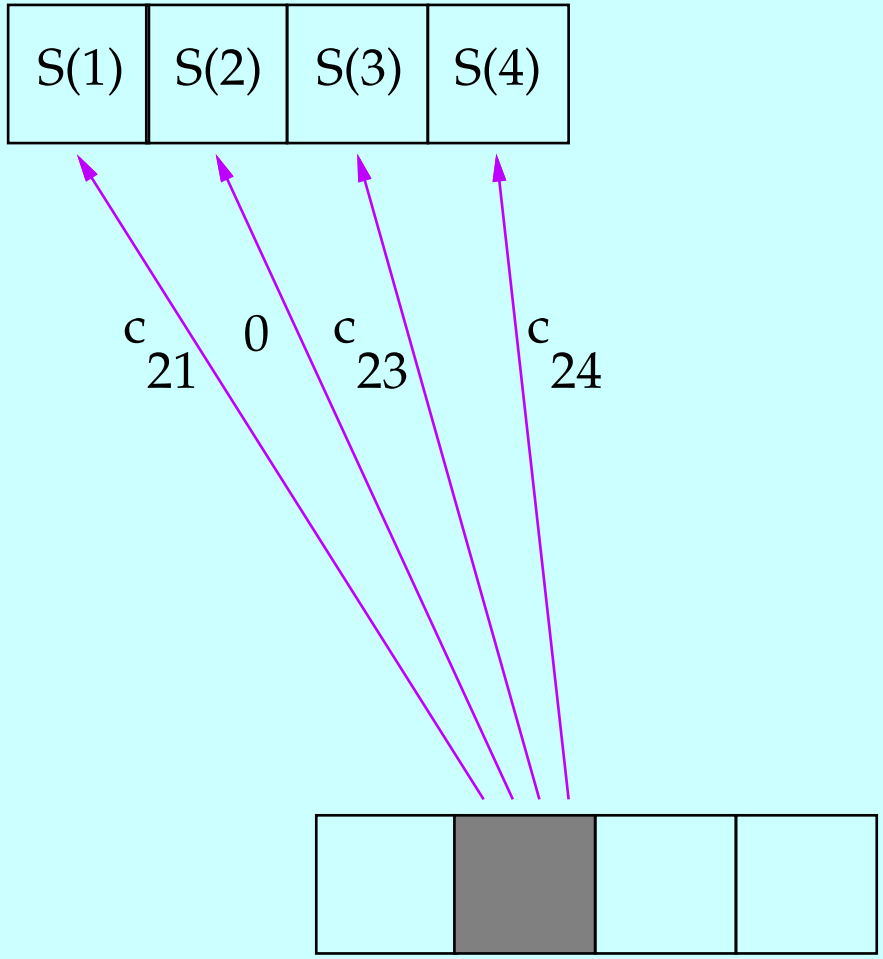
Inferring ancestor at the root of the tree



Cost matrix:

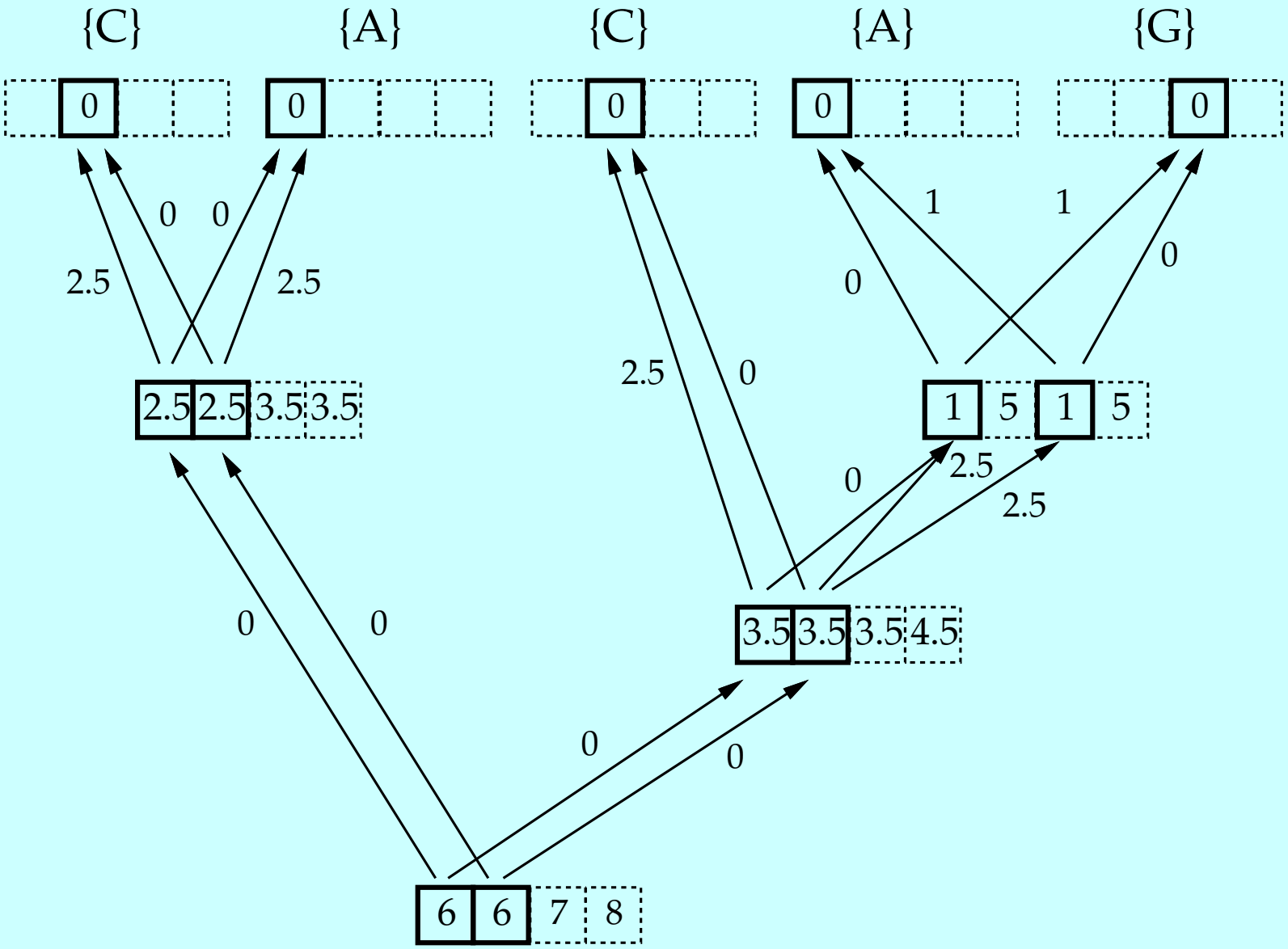
from \ to	A	C	G	T
A	0	2.5	1	2.5
C	2.5	0	2.5	1
G	1	2.5	0	2.5
T	2.5	1	2.5	0

Parsimony reconstruction of ancestral states

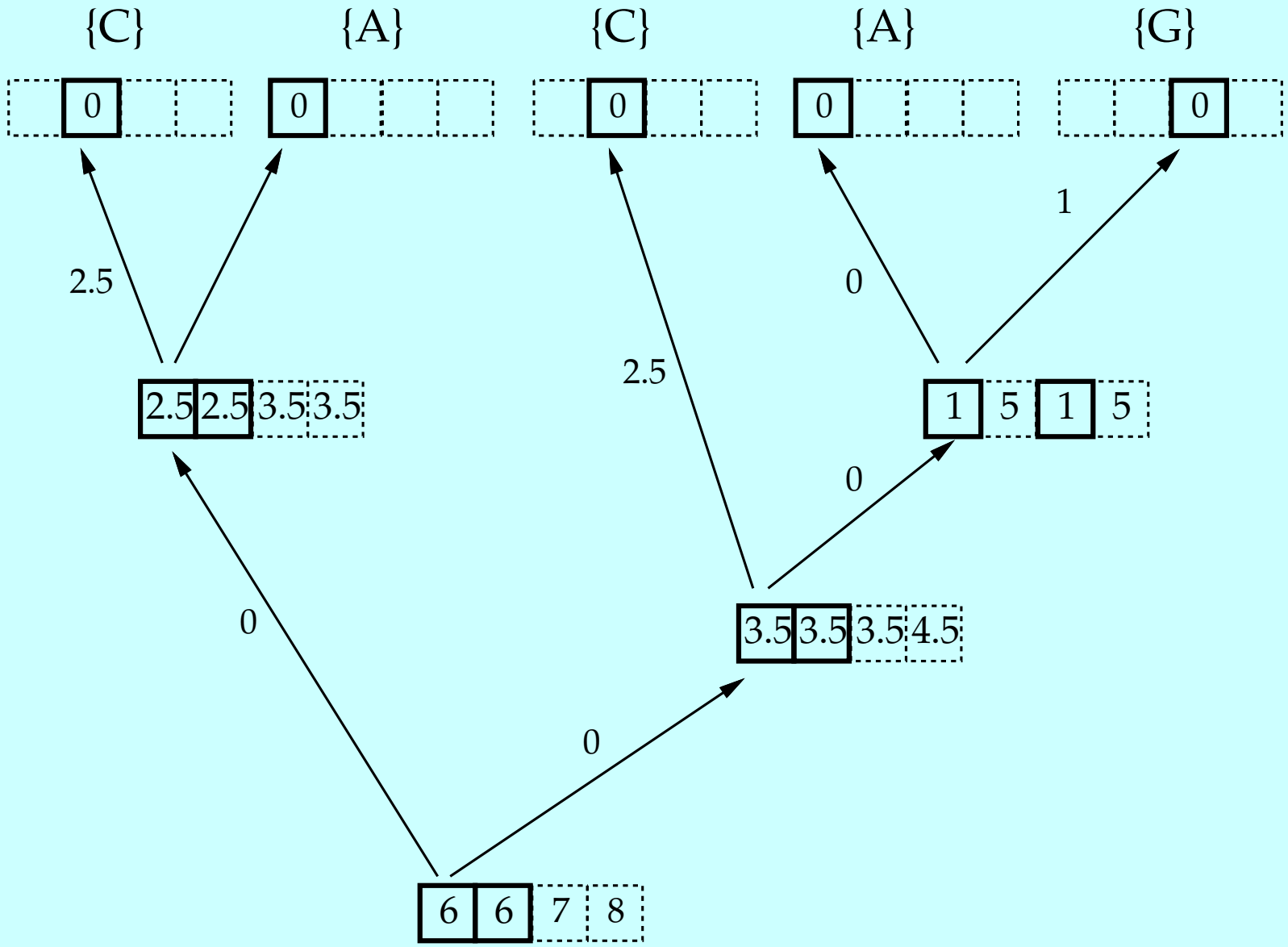


Given the reconstruction in an ancestor, choosing the most parsimonious one in one of its descendants.

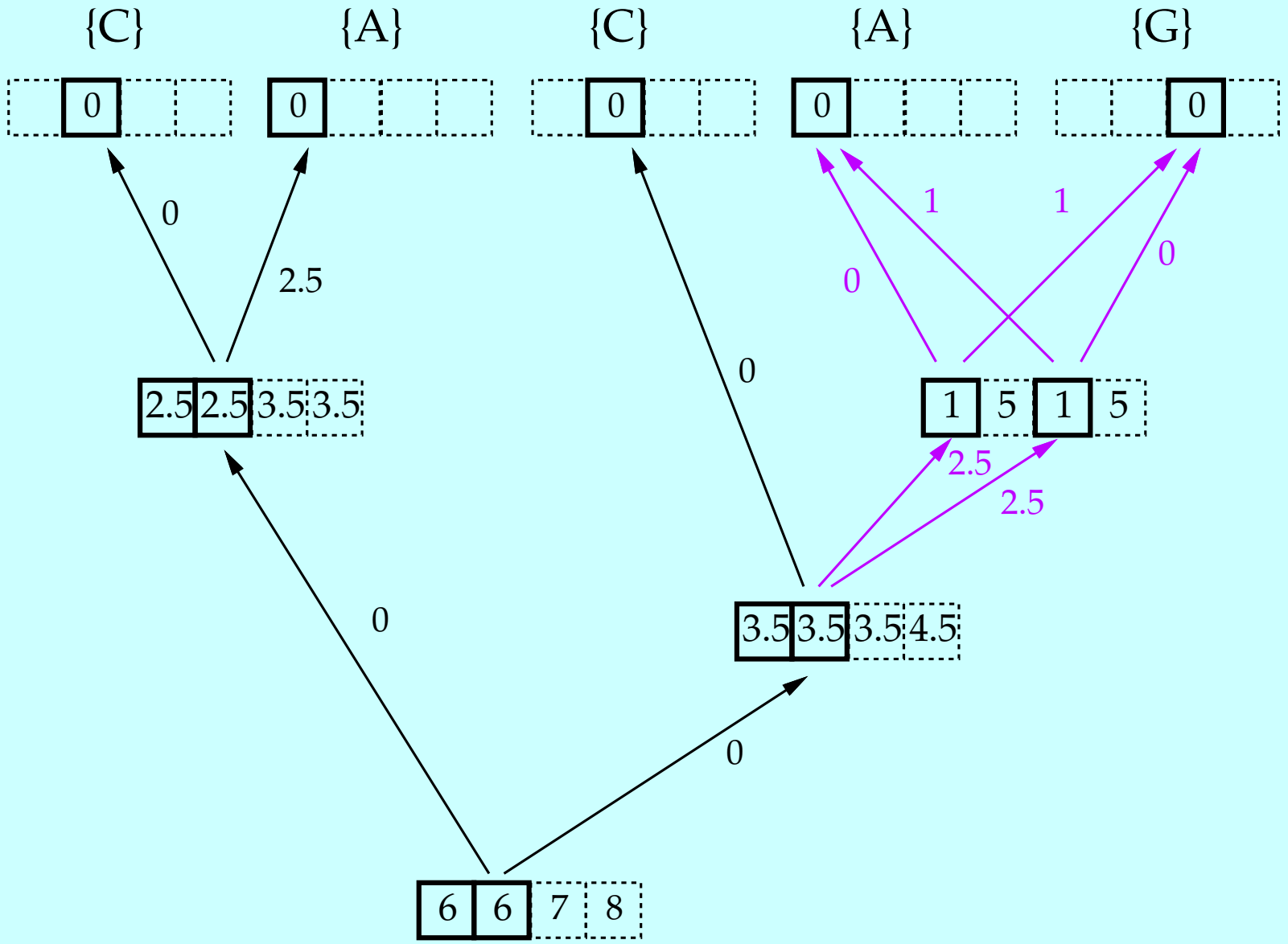
Ancestral states in the example



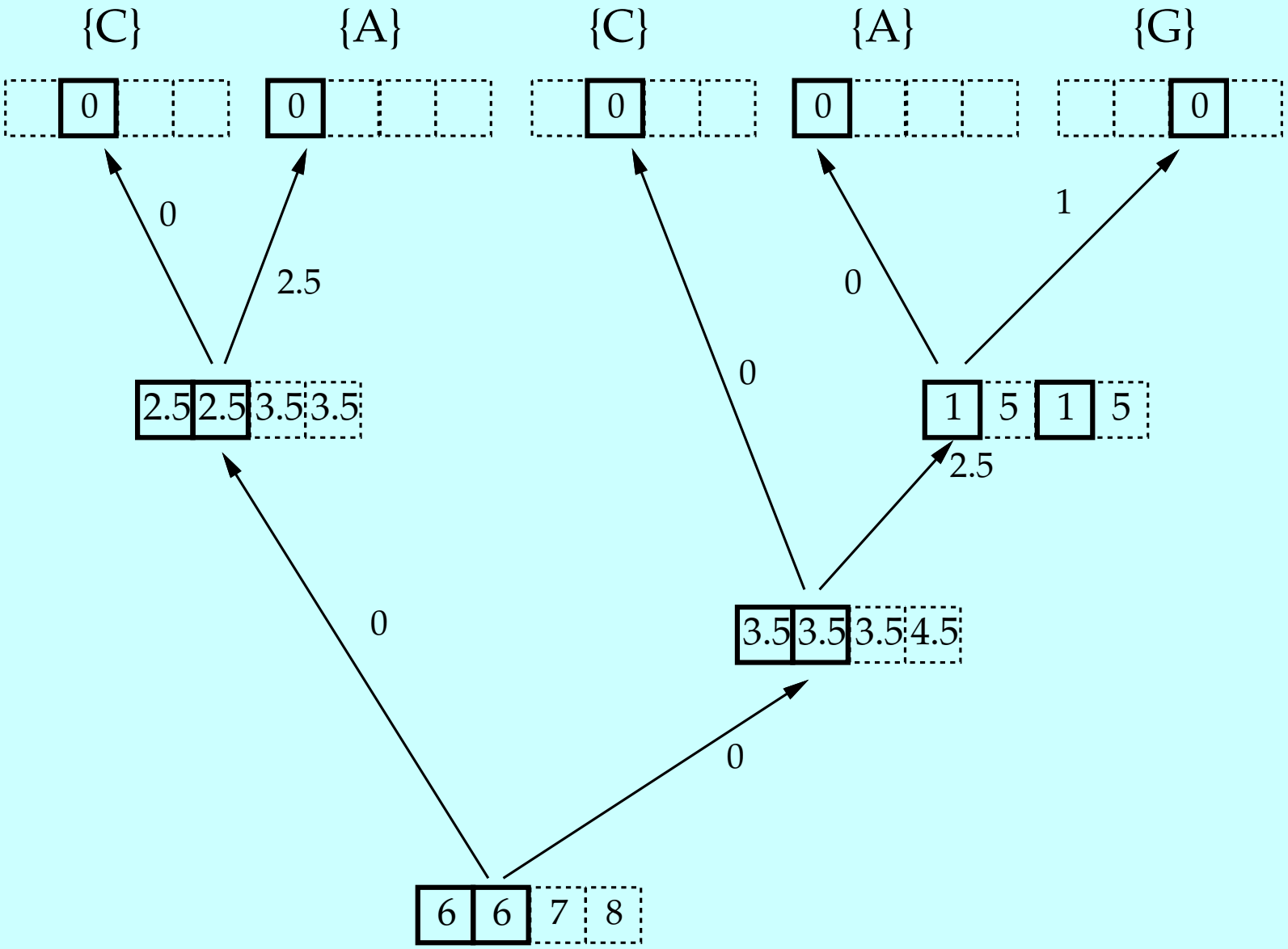
One reconstruction



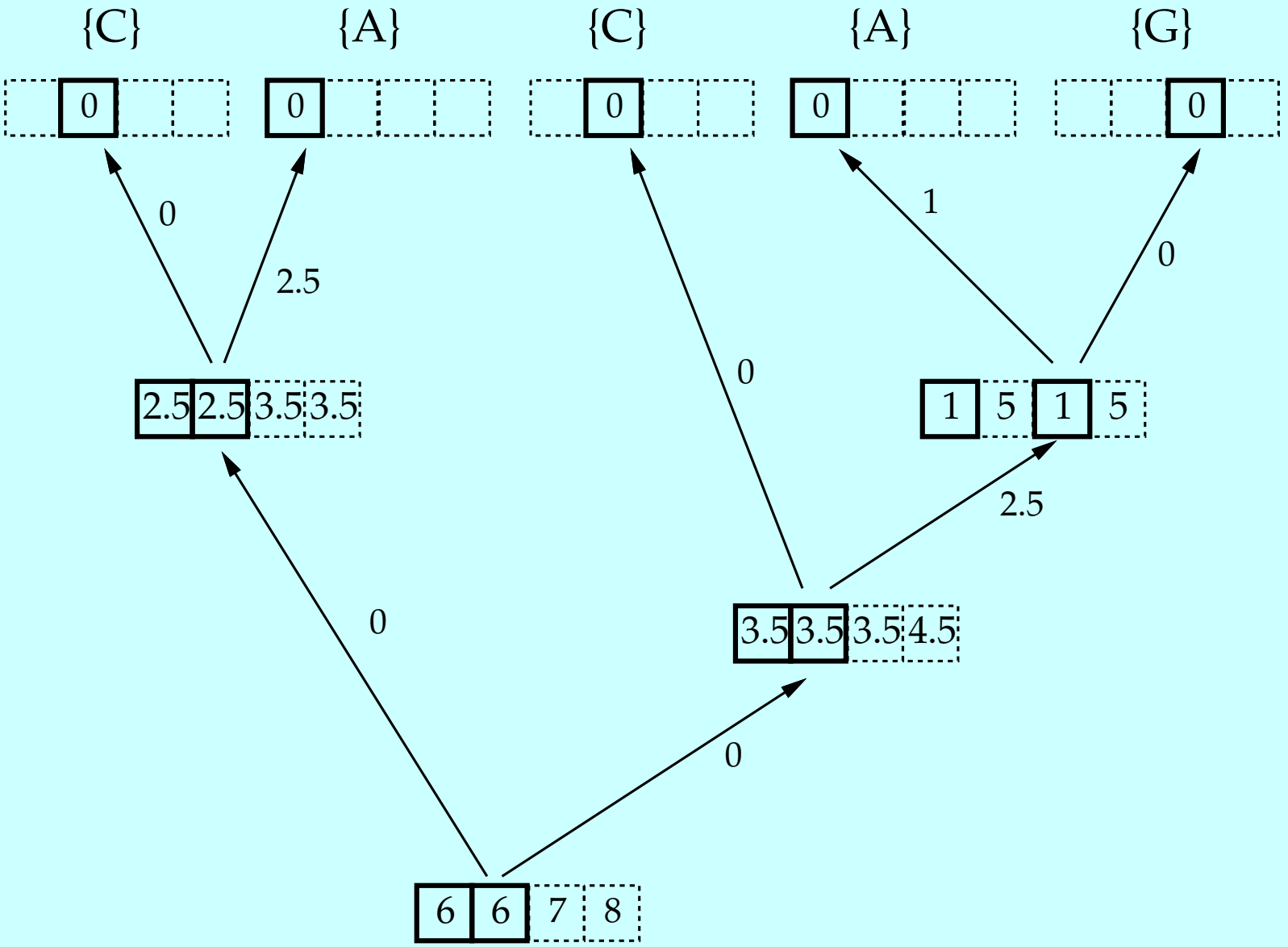
For the other choice, two possibilities



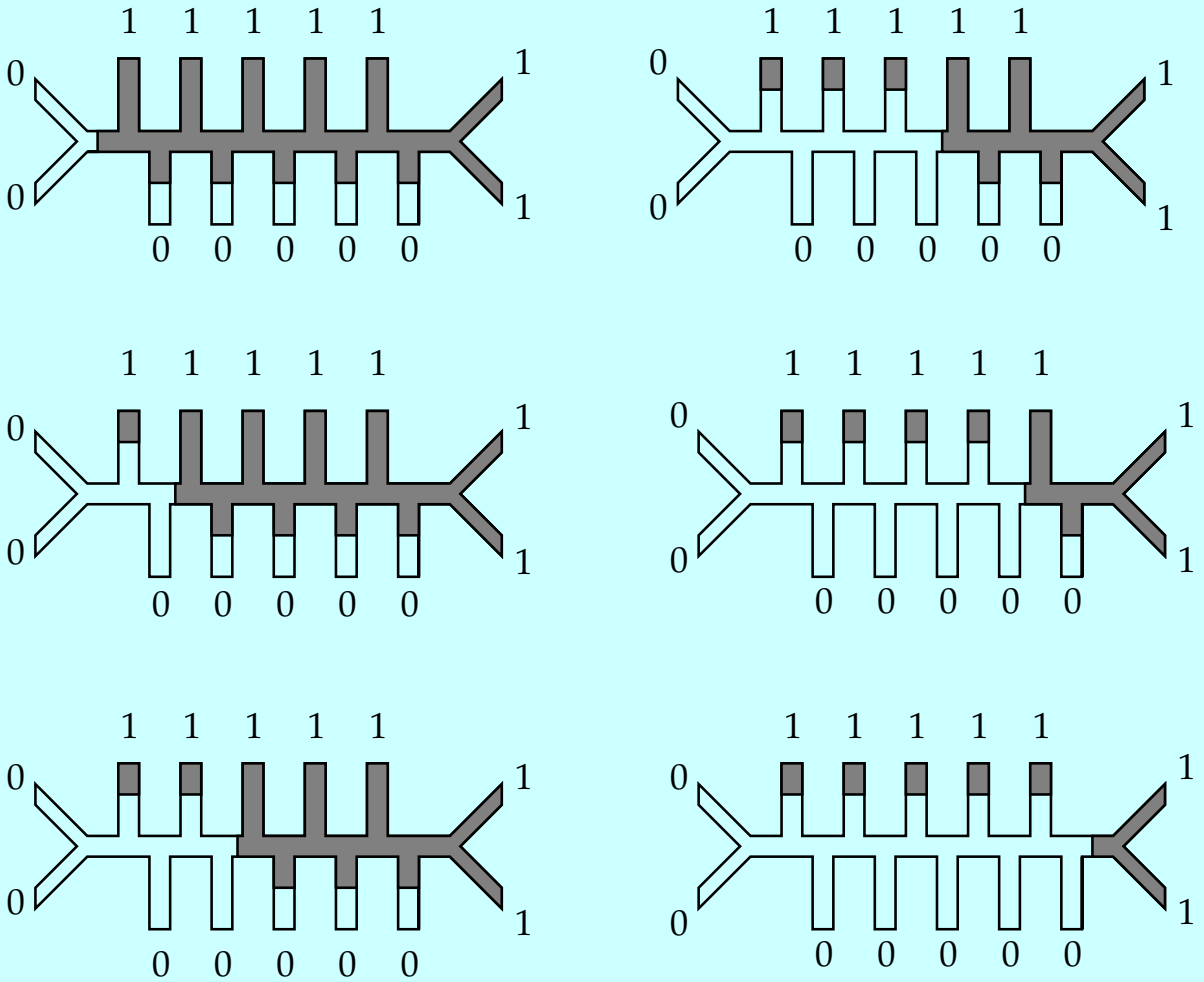
One of these



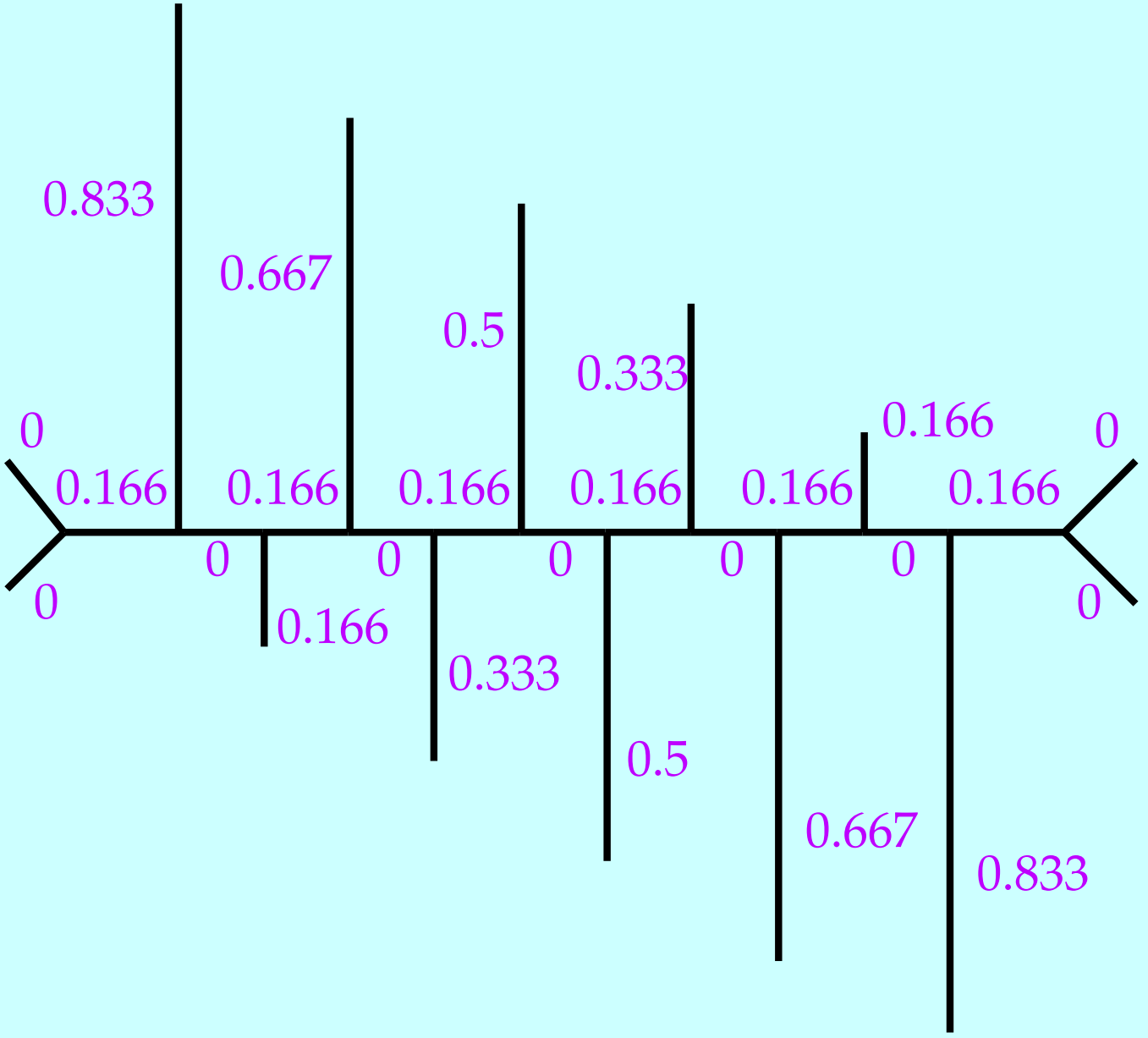
The other one



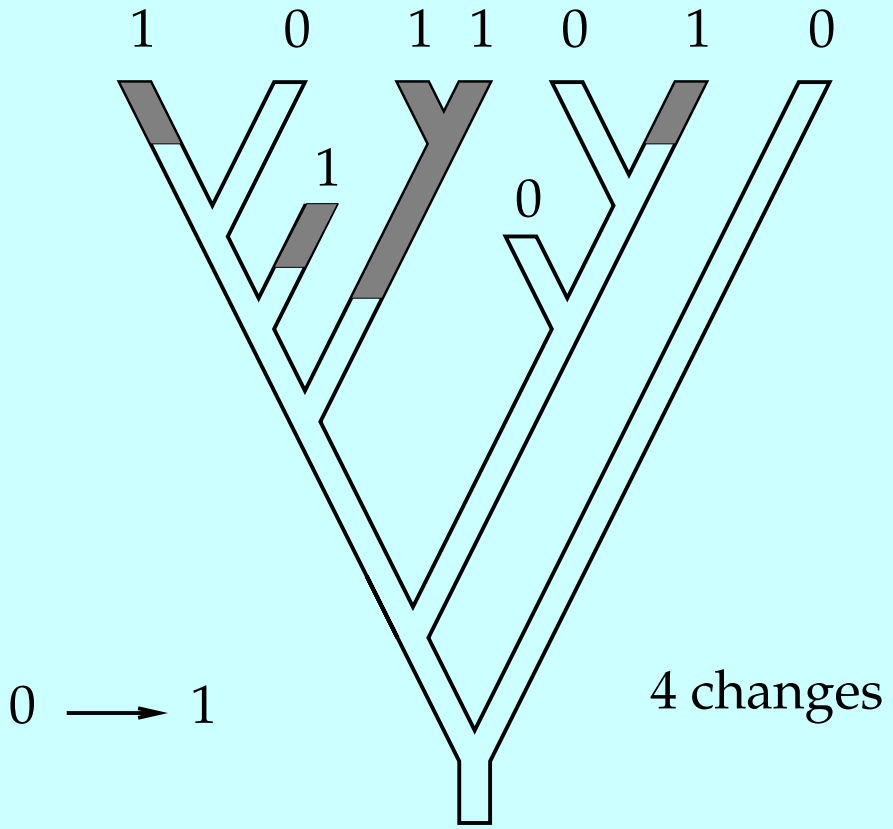
There can be multiple tied reconstructions



Average branch lengths over all reconstructions

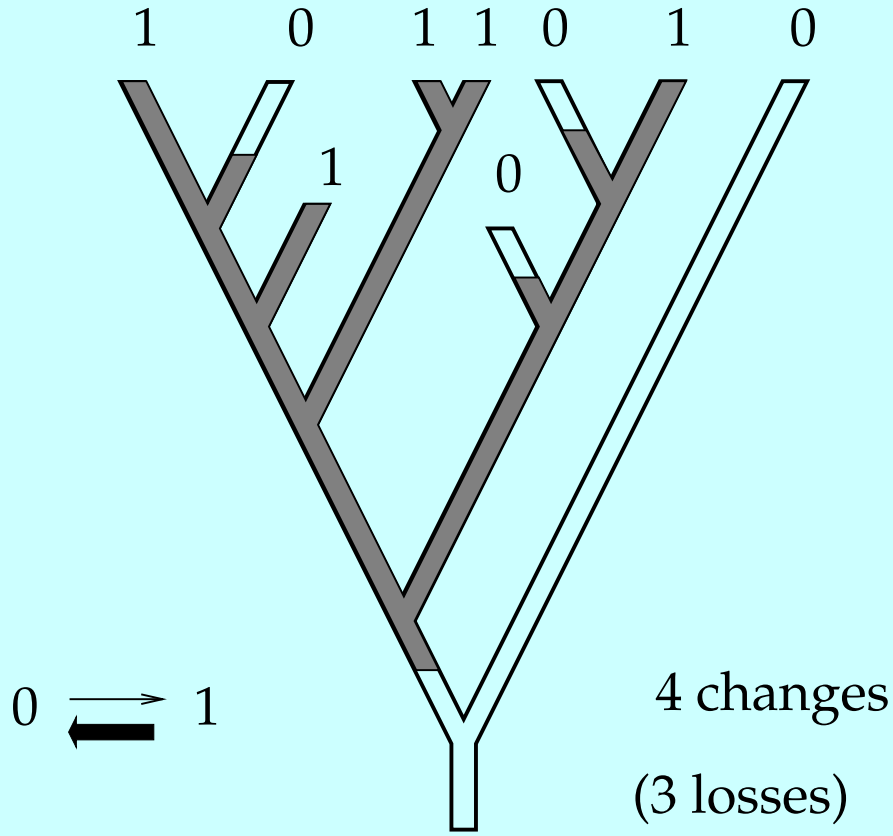


Camin-Sokal parsimony



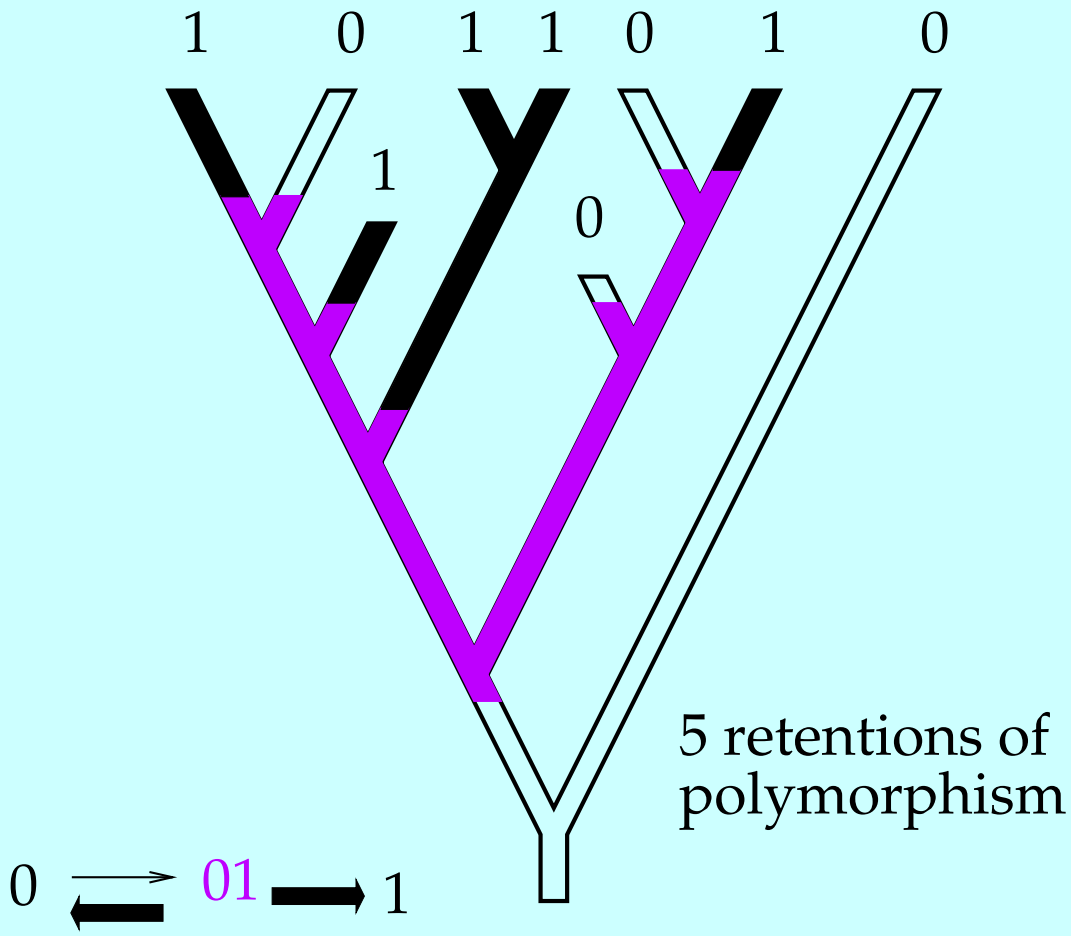
Unidirectional change. Easy to reconstruct ancestors and count states. This scheme is of importance in comparative genomics, with 1 = presence of a deletion.

Dollo parsimony



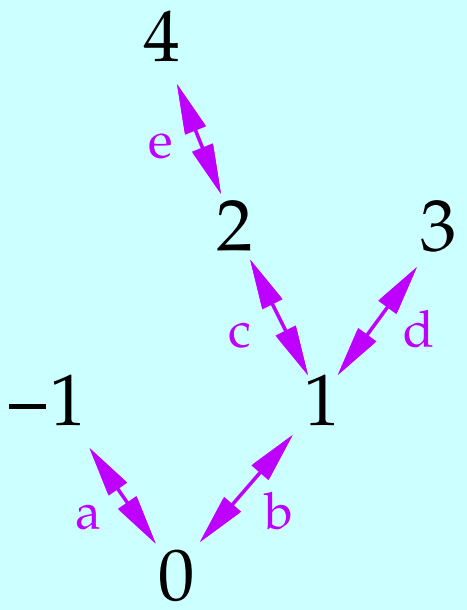
Assumes that it is much more difficult to gain state 1 than to lose it. Has been used to model gain and loss of restriction sites.

Polymorphism parsimony



Note that in this case the retention (not the loss) of the polymorphic state (01) is considered unlikely and is penalized.

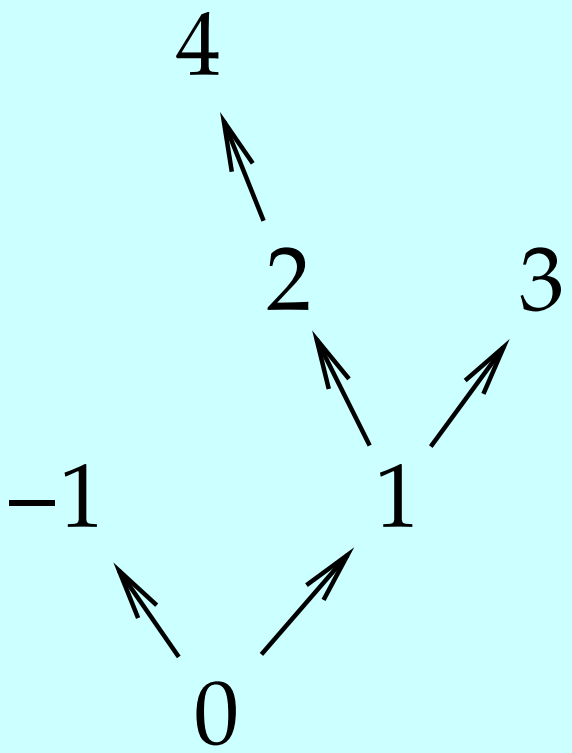
nonadditive binary coding

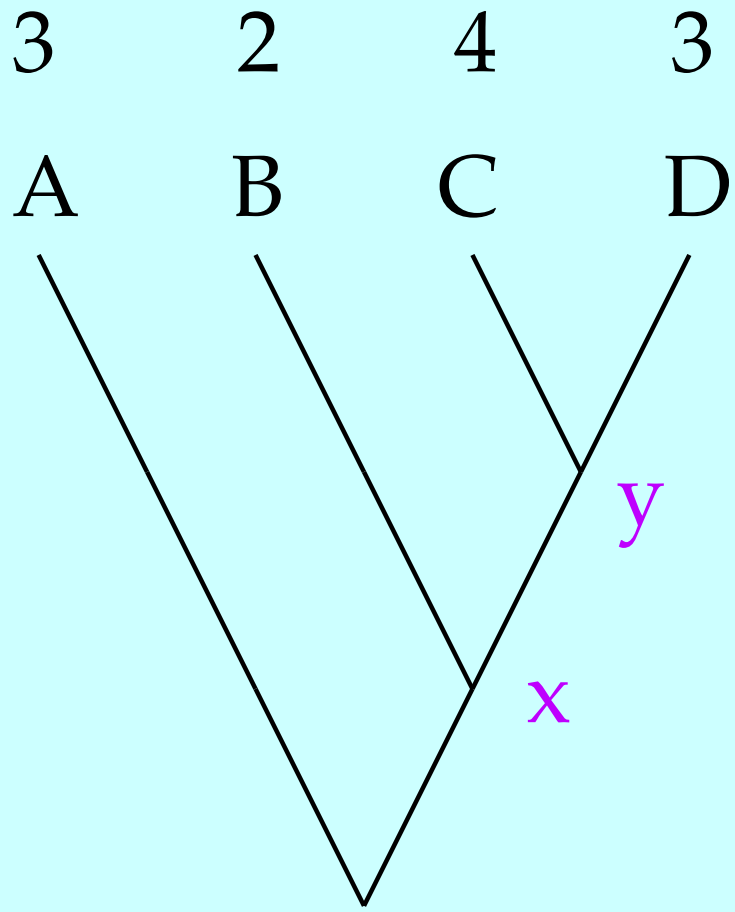


original	new characters a	b	c	d	e
–1	1	0	0	0	0
0	0	0	0	0	0
1	0	1	0	0	0
2	0	1	1	0	0
3	0	1	0	1	0
4	0	1	1	0	1

Using multiple 0/1 characters to construct a case with the same parsimony score on all trees as a single multistate character with a "character state tree"

Dollo parsimony – a paradox





With a character-state tree, can you always find a Dollo parsimony reconstruction that has only one origin of state 2 ?