

Bioinformatics I

lecture 15

- Tree properties
- Tree comparison
- Tree evaluation
- Tree significance (bootstrapping)

Properties of patristic distances

METRIC DISTANCES between any two or three taxa (a, b, and c) have the following properties:

Property 1: $d(a, b) \geq 0$

Property 2: $d(a, b) = d(b, a)$

Property 3: $d(a, b) = 0$ if and only if $a = b$

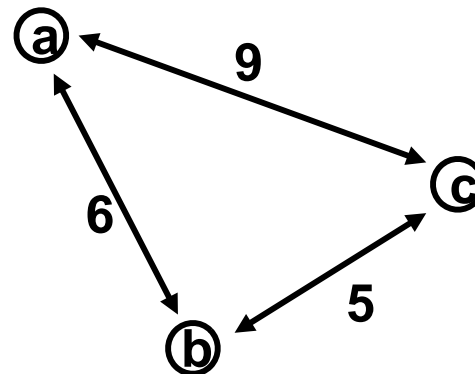
Property 4: $d(a, c) \leq d(a, b) + d(b, c)$

Non-negativity

Symmetry

Distinctness

Triangle inequality



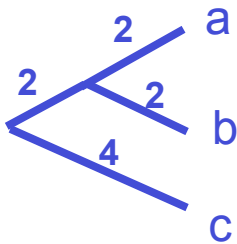
triangle inequality

A special distance metric..

ULTRAMETRIC DISTANCES

....must satisfy the previous four conditions, plus:

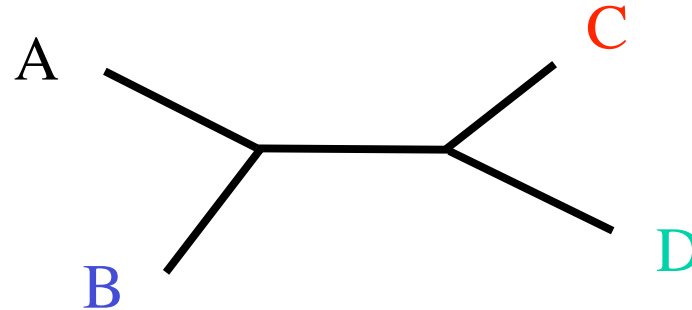
Property 5 *The distances from any branch point to the taxa in the clade defined by that branch point are equal.*



If distances are *ultrametric*, then the sequences are evolving in a perfectly **clock-like manner**. So any two sequences always have the same distance to their common ancestor.

One more property of patristic distances

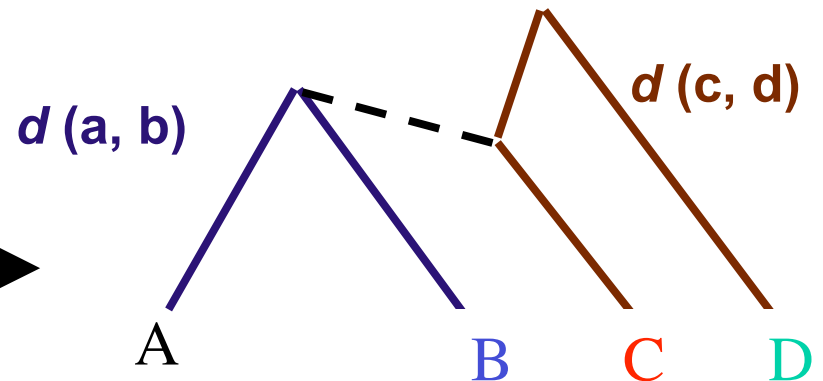
Additivity



Property 6: Example: if (a,b) are nearest neighbors,
 $d(a, b) + d(c, d) \leq \text{maximum} [d(a, c) + d(b, d), d(a, d) + d(b, c)]$

For distances to fit into an evolutionary tree, they must be additive. Estimated distances often fall short of these criteria, and thus can fail to produce correct evolutionary trees.

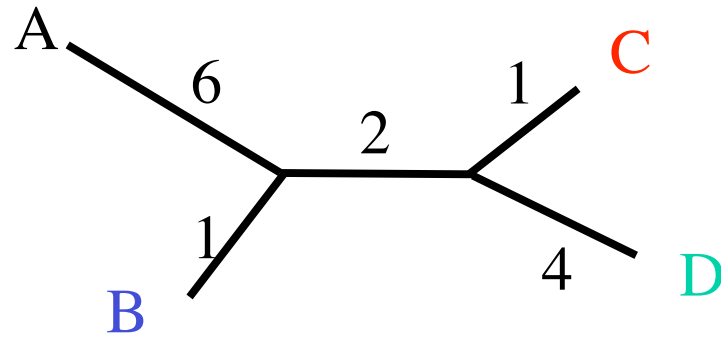
A lineage that violates additivity implicitly goes *backwards in time*.



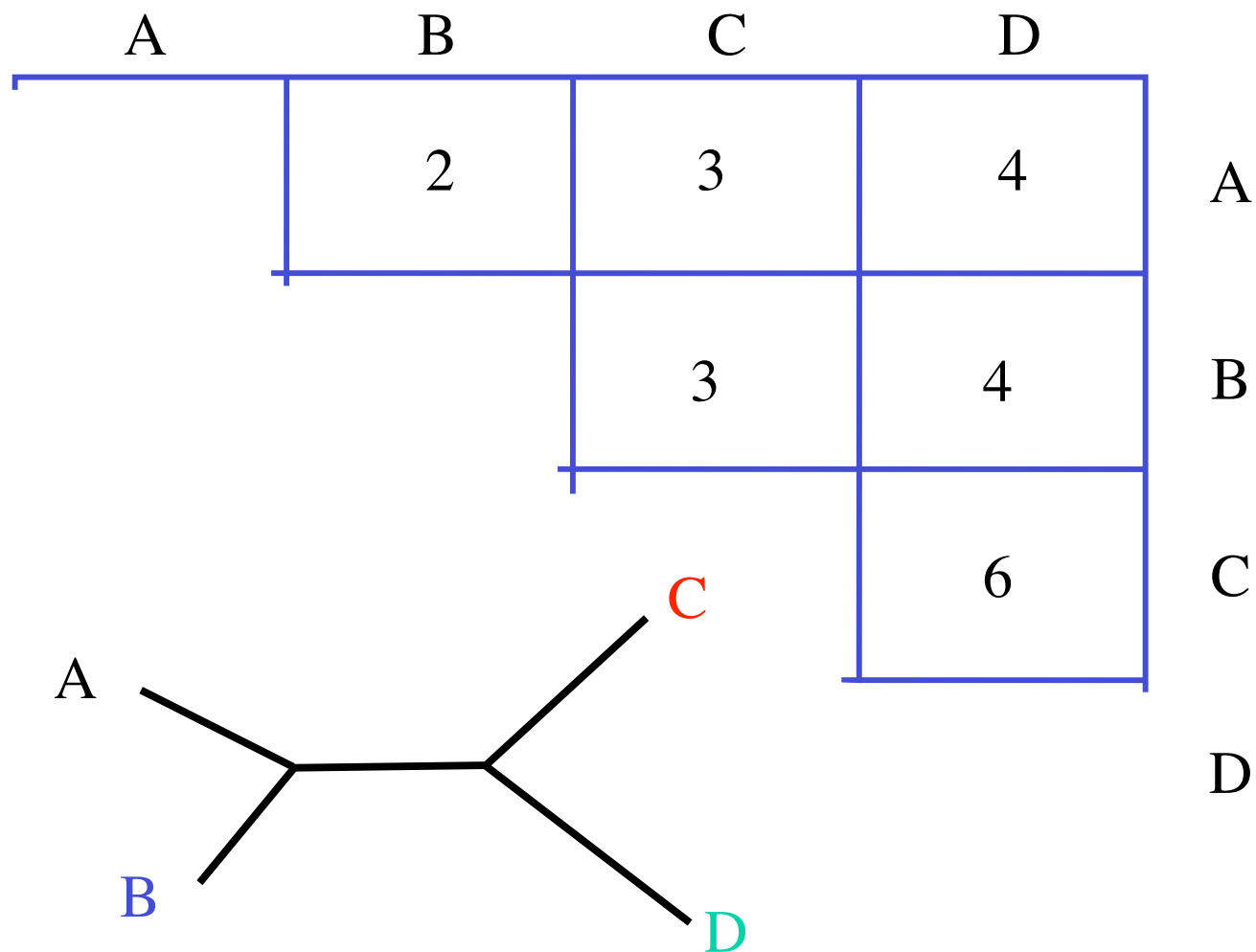
What's wrong with these distances?

	A	B	C	D
A	0	3	5	7
B	3	0	1	4
C	5	1	0	9
D	7	4	9	0

What's wrong with this tree and these branch lengths?



What's wrong with these distances
on this tree?



Tree comparison

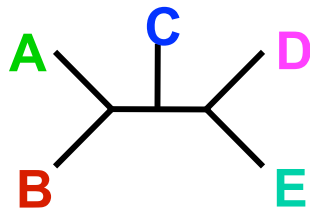
- If two trees have the same topology, compare the branch lengths. $S = \sum_{i>j} w_{ij} (d_{ij} - \Delta_{ij})^m$

- If two trees have different topologies, calculate the **symmetric distance**.

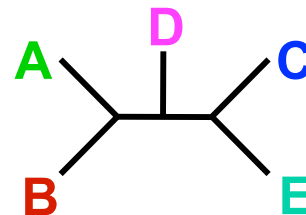
D = number of “splits” that are different.

A “split” is a bisection of the tree at an internal (not leaf) lineage.

A split is different if the taxa on either side of the split are different.



Splits
AB, CDE
ABC, DE

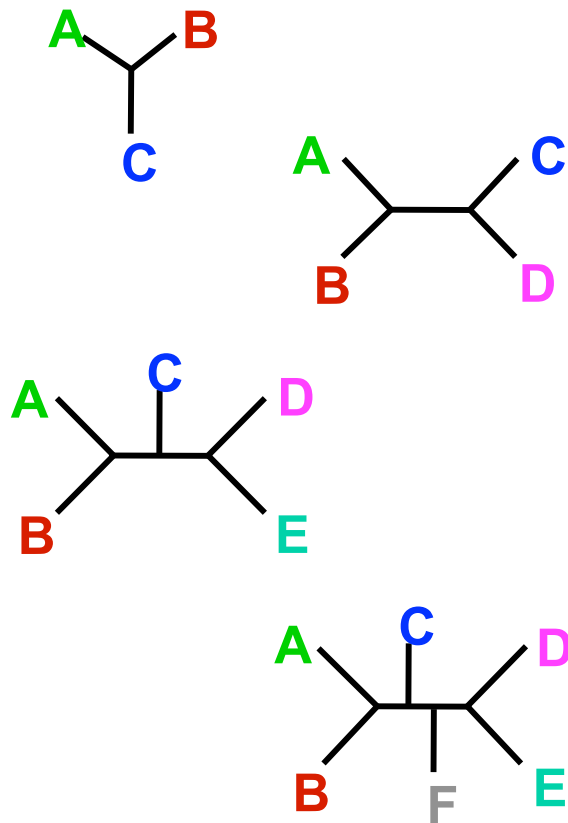


Splits
AB, CDE
ABD, CE

Tree evaluation

- A tree topology may be evaluated relative to a set of distances. To select the tree that best fits the data we need...
 - (1)** A way to score each tree.
 - (a) Parsimony
 - (b) Fitch-Margoliash
 - (c) Least-squares
 - (d) Maximum likelihood
 - (2)** A way to generate trees.
 - (a) Nearest neighbor interchange (NNI)
 - (b) Subtree pruning and re-grafting (SPR)
 - (c) Tree bisection and reconnection (TBR)
 - (d) Exhaustive??

Explosive tree growth



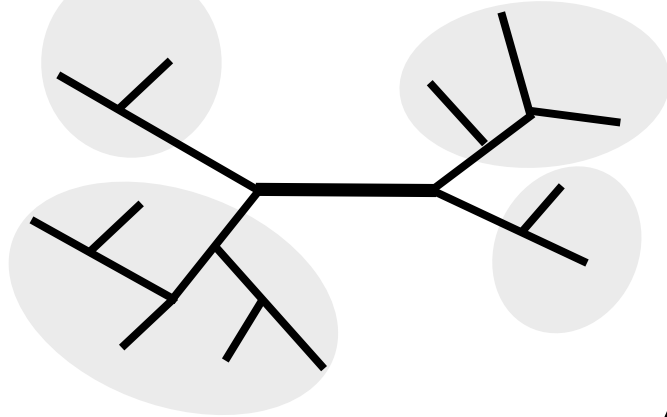
# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
·	·
·	·
·	·
·	·
·	·
30	≈ 3.58 × 10 ³⁶

Exhaustive search is possible only for small numbers of taxa.

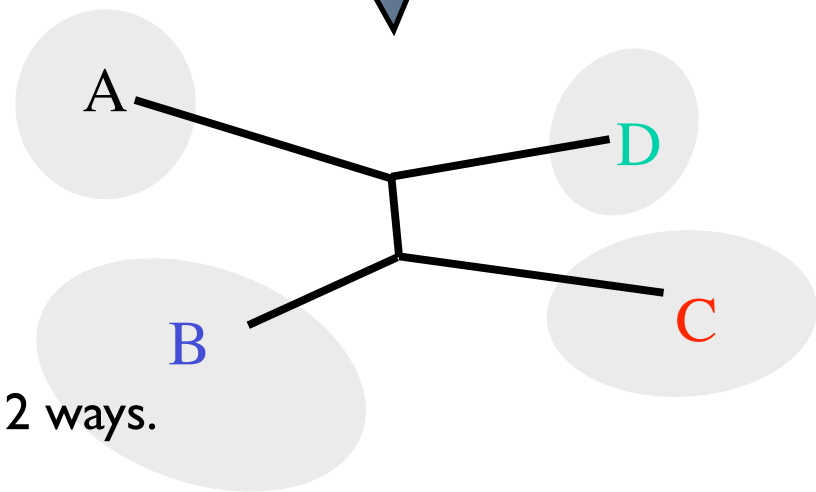
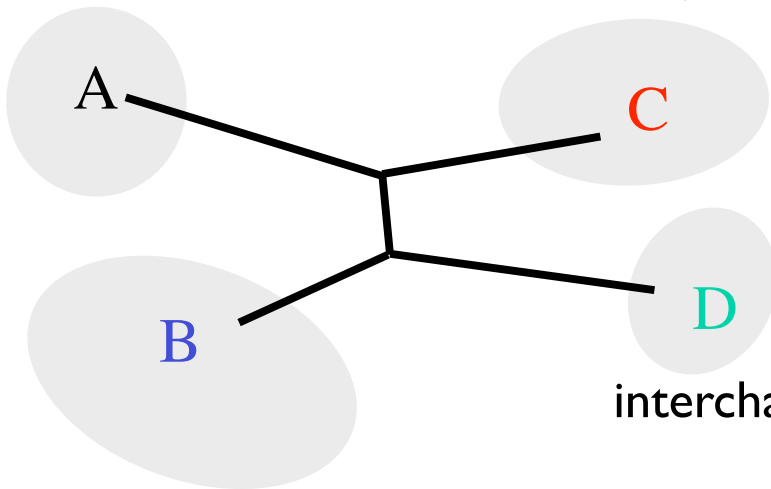
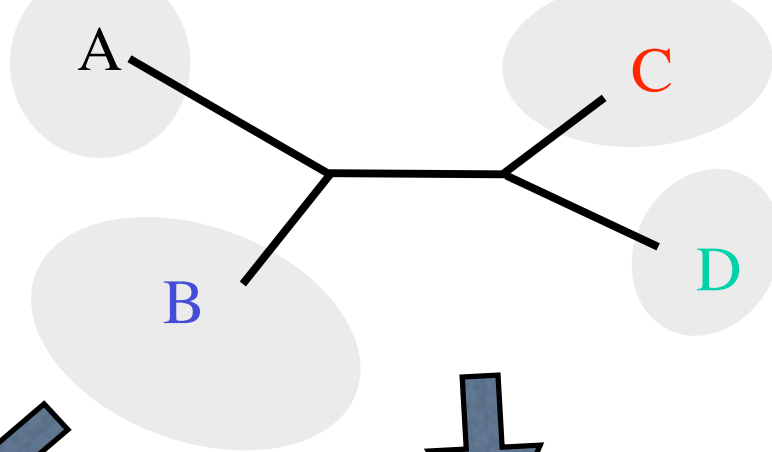
$$(2N-5)! / [2^{N-3} * (N-3)!] = \# \text{ unrooted trees for } N \text{ taxa}$$

Nearest neighbor interchange

Choose internal branch (lineage)



Condense to 4 groups (clades)

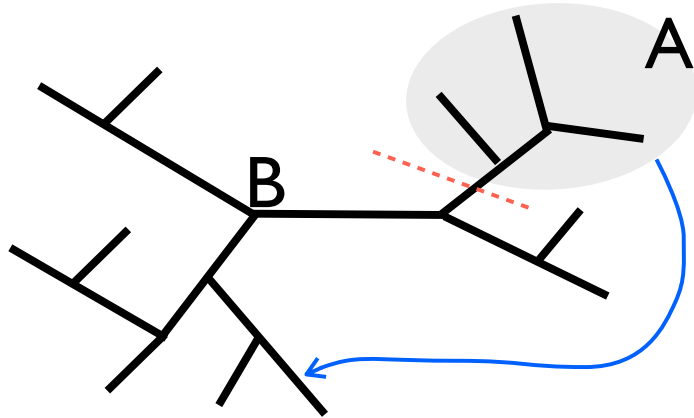


interchange, 2 ways.

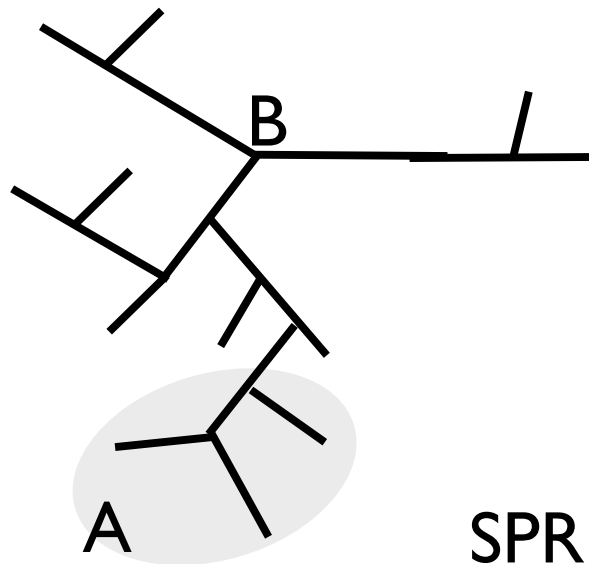
NNI makes small changes in the tree “Local search”

Subtree pruning and re-grafting

Choose internal branch (lineage)



Pick a branch.
Split the tree. (part A,
part B)

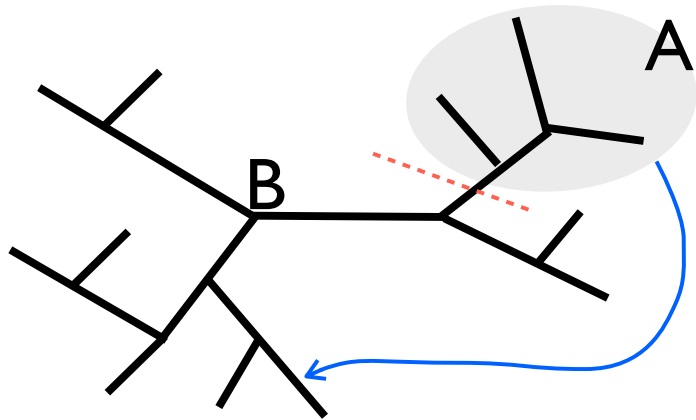


Pick a branch on Part
B.
Graft the cut point of
Part A to that branch.

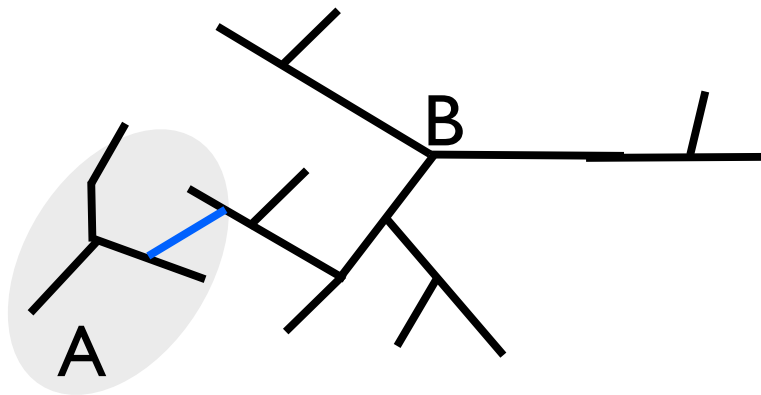
SPR makes large changes in the tree.

Tree bisection and reconnection

Choose internal branch (lineage)



Pick a branch.
Split the tree. (part A,
part B)



Randomly select one
lineage in A, one
lineage in B.
Connect them.

TBR makes even larger changes in the tree.

Scoring trees based on parsimony

- Requires a MSA
- Considers the characters.
- May be done using non-character, non-metric information, such as phenotypes.
- Assumes minimum evolution.

Scoring trees based on distances

- Compare sequence distances (J-C corrected) to patristic distances.
- Does not consider the characters.
- May not be done using non-character, non-metric information, such as phenotypes.
- No minimum evolution assumption. Maximizes correlation.
- Ways to get patristic distances:
 - (1) UPGMA (rooted tree. ultrametric distances.)
 - (2) Fitch-Margoliash (unrooted or rooted)
 - (3) Least-squares (unrooted or rooted)

$$S = \sum_{i>j} w_{ij} (d_{ij} - \Delta_{ij})^m$$

weights. may be a function of d_{ij}

sequence distances

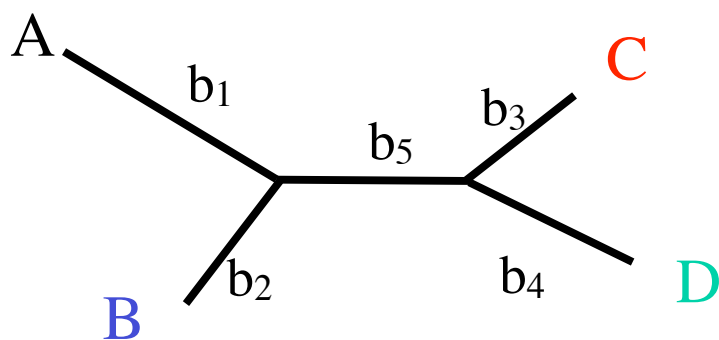
tree (patristic) distances

may be anything. typically 2

Least squares solution for tree branch lengths

$$S = \sum_{i>j} w_{ij} (d_{ij} - \Delta_{ij})^m$$

$$\begin{aligned} d_{AB} &= b_1 + b_2 \\ d_{AC} &= b_1 + b_3 + b_5 \\ &\text{etc} \end{aligned}$$



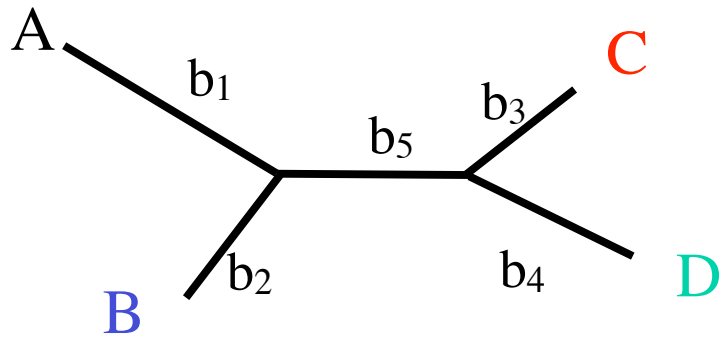
$$Tb = d$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{matrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{matrix} = \begin{matrix} d_{AB} \\ d_{AC} \\ d_{AD} \\ d_{BC} \\ d_{BD} \\ d_{CD} \end{matrix}$$

The T matrix expresses the branches needed to sum the tree distance. Branch distances **b** are found by squaring and inverting the T matrix and multiplying by the sequence distances **d**. New tree distances Δ are found by summing **b**, using T.

$$b = (T^T T)^{-1} (T^T d)$$

Least squares : solve for \mathbf{b}



$$T\mathbf{b} = \mathbf{d}$$

1	1	0	0	0	b_1	$=$	d_{AB}	$=$	3
1	0	1	0	1	b_2	$=$	d_{AC}	$=$	5
1	0	0	1	1	b_3	$=$	d_{AD}	$=$	7
0	1	1	0	1	b_4	$=$	d_{BC}	$=$	4
0	1	0	1	1	b_5	$=$	d_{BD}	$=$	6
0	0	1	1	0		$=$	d_{CD}	$=$	3

$$\mathbf{b} = (T^T T)^{-1} (T^T \mathbf{d})$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}
 \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}
 =
 \begin{bmatrix} 3 & 1 & 1 & 1 & 2 \\ 1 & 3 & 1 & 1 & 2 \\ 1 & 1 & 3 & 1 & 2 \\ 1 & 1 & 1 & 3 & 2 \\ 2 & 2 & 2 & 2 & 4 \end{bmatrix}$$

Invert this to get $(T^T T)^{-1}$

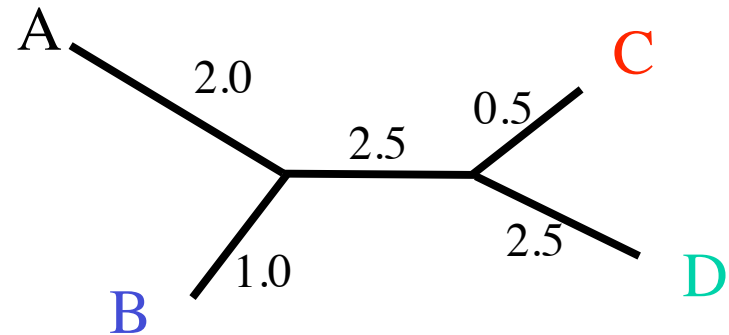
$$T^T T = \begin{bmatrix} 3 & 1 & 1 & 1 & 2 \\ 1 & 3 & 1 & 1 & 2 \\ 1 & 1 & 3 & 1 & 2 \\ 1 & 1 & 1 & 3 & 2 \\ 2 & 2 & 2 & 2 & 3 \end{bmatrix} \quad (T^T T)^{-1} = \begin{bmatrix} 1/2 & 0 & 0 & 0 & -1/4 \\ 0 & 1/2 & 0 & 0 & -1/4 \\ 0 & 0 & 1/2 & 0 & -1/4 \\ 0 & 0 & 0 & 1/2 & -1/4 \\ -1/4 & -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix}$$

$$(T^T d) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{matrix} 3 \\ 5 \\ 7 \\ 4 \\ 6 \\ 3 \end{matrix} = \begin{matrix} 15 \\ 13 \\ 12 \\ 16 \\ 22 \end{matrix}$$

$$b = (T^T T)^{-1} (T^T d) = \begin{bmatrix} 1/2 & 0 & 0 & 0 & -1/4 \\ 0 & 1/2 & 0 & 0 & -1/4 \\ 0 & 0 & 1/2 & 0 & -1/4 \\ 0 & 0 & 0 & 1/2 & -1/4 \\ -1/4 & -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix} \begin{matrix} 15 \\ 13 \\ 12 \\ 16 \\ 22 \end{matrix} = \begin{matrix} 2.0 \\ 1.0 \\ 0.5 \\ 2.5 \\ 2.5 \end{matrix}$$

*thanks to http://wims.unice.fr/wims/en_home.html

$$\begin{matrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{matrix} = (T^T T)^{-1} (T^T d) = \begin{matrix} 2.0 \\ 1.0 \\ 0.5 \\ 2.5 \\ 2.5 \end{matrix}$$



Sequence distances

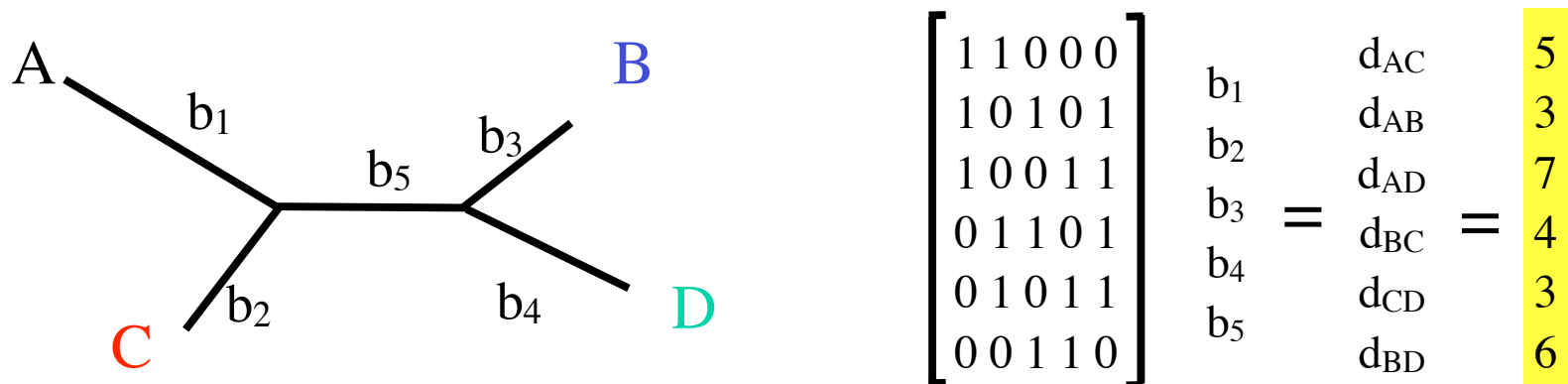
$$\begin{matrix} d_{AB} \\ d_{AC} \\ d_{AD} \\ d_{BC} \\ d_{BD} \\ d_{CD} \end{matrix} = \begin{matrix} 3 \\ 5 \\ 7 \\ 4 \\ 6 \\ 3 \end{matrix}$$

Least squares patristic distances

$$\begin{matrix} \Delta_{AB} \\ \Delta_{AC} \\ \Delta_{AD} \\ \Delta_{BC} \\ \Delta_{BD} \\ \Delta_{CD} \end{matrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{matrix} 2.0 \\ 1.0 \\ 0.5 \\ 2.5 \\ 2.5 \end{matrix} = \begin{matrix} 3.0 \\ 5.0 \\ 7.0 \\ 4.0 \\ 6.0 \\ 3.0 \end{matrix}$$

$$S = \sum_{i>j} (d_{ij} - \Delta_{ij})^2 = 0.0$$

Least squares distances for the wrong tree



$$\mathbf{b} = (\mathbf{T}^T \mathbf{T})^{-1} (\mathbf{T}^T \mathbf{d}) =
 \begin{bmatrix} 1/2 & 0 & 0 & 0 & -1/4 \\ 0 & 1/2 & 0 & 0 & -1/4 \\ 0 & 0 & 1/2 & 0 & -1/4 \\ 0 & 0 & 0 & 1/2 & -1/4 \\ -1/4 & -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix}
 \begin{bmatrix} 5 \\ 3 \\ 7 \\ 4 \\ 3 \\ 6 \end{bmatrix}
 =
 \begin{bmatrix} 1/2 & 0 & 0 & 0 & -1/4 \\ 0 & 1/2 & 0 & 0 & -1/4 \\ 0 & 0 & 1/2 & 0 & -1/4 \\ 0 & 0 & 0 & 1/2 & -1/4 \\ -1/4 & -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix}
 \begin{bmatrix} 15 \\ 12 \\ 13 \\ 16 \\ 17 \end{bmatrix}
 =
 \begin{bmatrix} 3.25 \\ 1.75 \\ 2.25 \\ 3.75 \\ -1.25 \end{bmatrix}$$

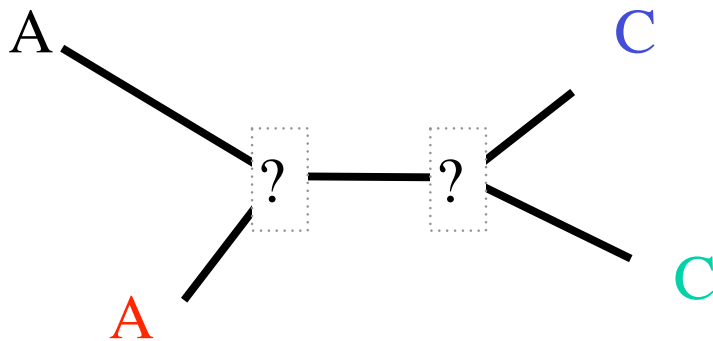
$$\begin{matrix} \Delta_{AC} \\ \Delta_{AB} \\ \Delta_{AD} \\ \Delta_{BC} \\ \Delta_{CD} \\ \Delta_{BD} \end{matrix}
 =
 \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}
 \begin{matrix} 3.25 \\ 1.75 \\ 2.25 \\ 3.75 \\ 0.00 \\ 0.00 \end{matrix}
 =
 \begin{matrix} 5.0 \\ 5.5 \\ 7.0 \\ 4.0 \\ 5.5 \\ 6.0 \end{matrix}$$

$$S = \sum_{i>j} (d_{ij} - \Delta_{ij})^2 = 12.5$$

Least-squares produces a negative distance for b_5

Maximum likelihood

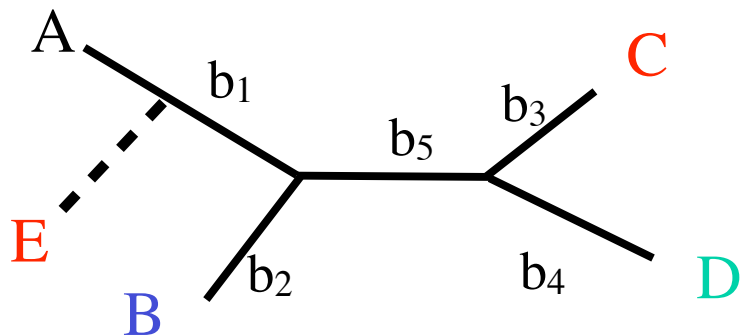
- [Given a tree with branch lengths and MSA, sum the probabilities of the branches over all possible ancestor bases (or amino acids).
 - [Probabilities may be J-C for nucleotide, or PAM or BLOSUM for protein.
- [May be used in combination with NNI, SPR, TBR to evaluate **tree topology**, with or without branch lengths (if no branch lengths, assume all branch length are equal).
- [May be used to produce ancestral sequences.



Sum the total likelihood of all branches over all possible bases at two ? positions.

Quartet puzzling

1. Identify all choose-4 subsets of MSA => “quartets”
2. For each quartet, find the best 4-taxa tree (3 possibilities) using least-squares, Fitch-Margoliash, Parsimony, or Maximum Likelihood.
3. Start with a randomly selected quartet, ((A,B),(C,D))
4. Choose one new taxon, E.
5. For every branch, try adding the E to the branch.
 - 5.1. For all quartets containing E and three of the other taxa, ask whether the split is correct. If not, add 1 to the penalty.
6. Add E to the branch with the lowest penalty.
7. Continue at step 4 with a new taxon.



	penalty for adding E to branch				
quartet	b1	b2	b3	b4	b5
AB,CD	-	-	-	-	-
AE,BC	0	1	1	1	1
AE,BD	0	1	1	1	1
AE,CD	0	0	1	1	0
BE,CD	0	0	1	1	0

QP can be repeated, starting with different quartet, to generate any number of trees.

Expert tree strategy

- Get sequence distances.
- **QP with ML** -- Use ML to choose topology for each quartet. Build trees from quartets.
- **NNI, SPR, TBR with L-S, F-M** -- Modify the tree, calculate branch lengths, calculate patristic distances, calculate difference distance score **S**.
- **Consensus tree** -- if more than one tree has same best score, merge branches.

“Boot strap analysis”

- A method to validate the significance of a phylogenetic tree, branchpoint by branchpoint.
- Requires a means to generate independent trees. (For example using distances generated from different regions of the mitochondrial genome or from random subsets of sequences or random subsets of columns.)
- Choose the representative tree as the ‘parent’. Calculate the following:

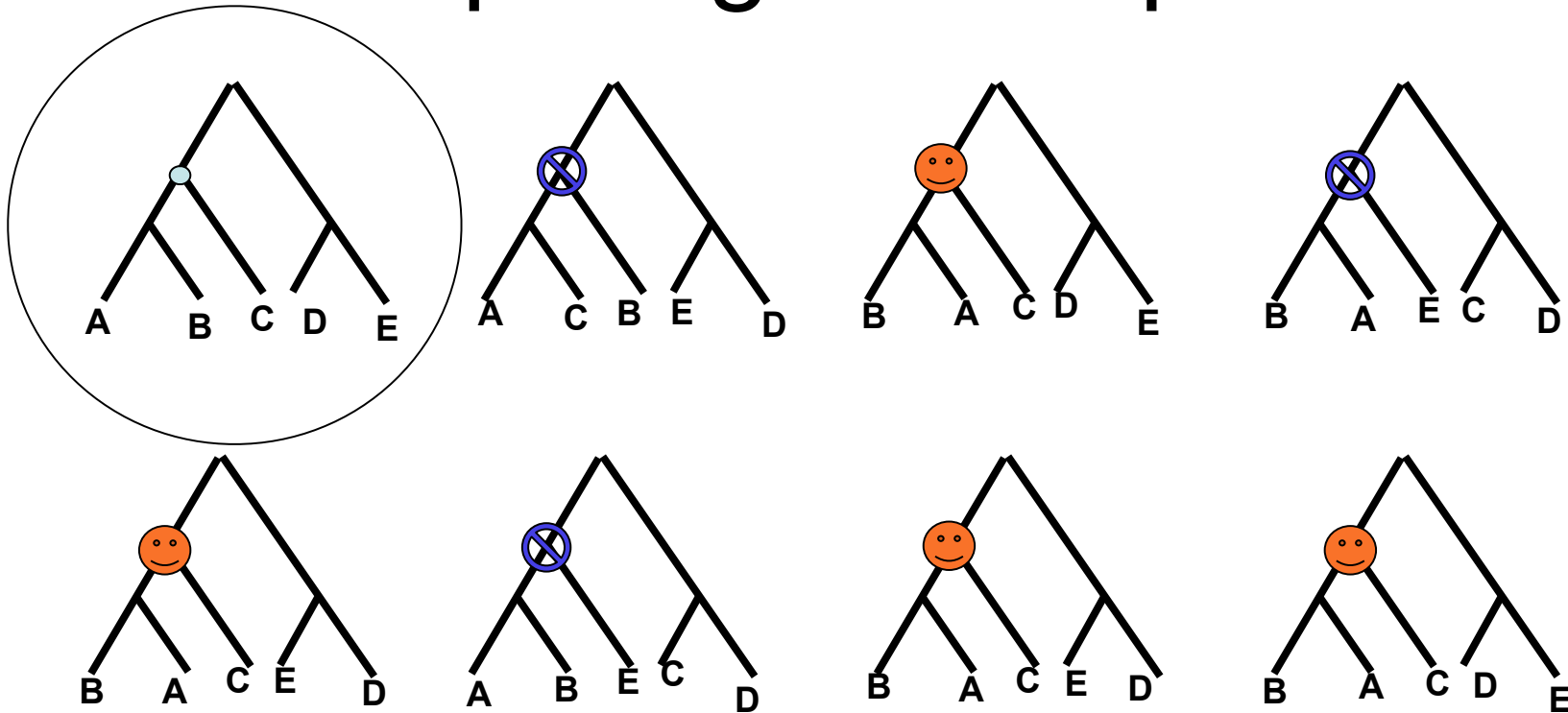
For each branchpoint in the parent tree,

For each tree, ask

Is there a branchpoint having the same subclade contents (i.e. same taxa, any order)

Bootstrap value = number of trees having the branchpoint / total trees.

Comparing branchpoints



$$\bullet = P((A,B),C) = 5/8$$

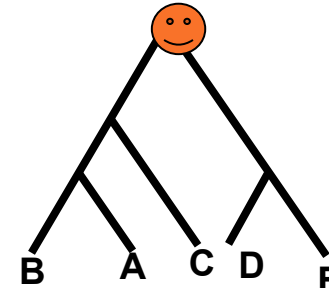
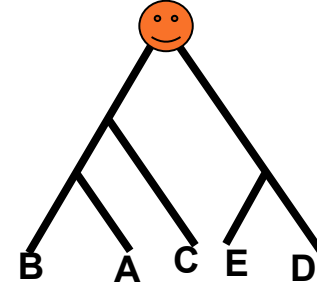
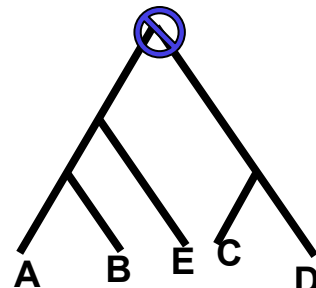
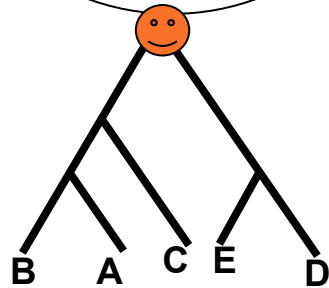
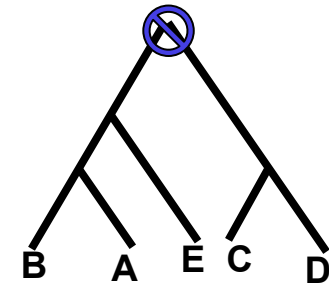
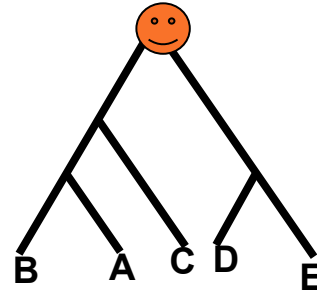
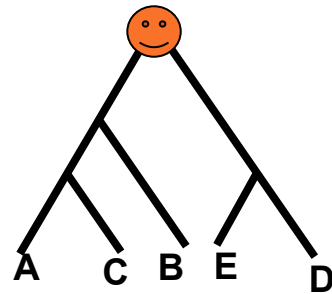
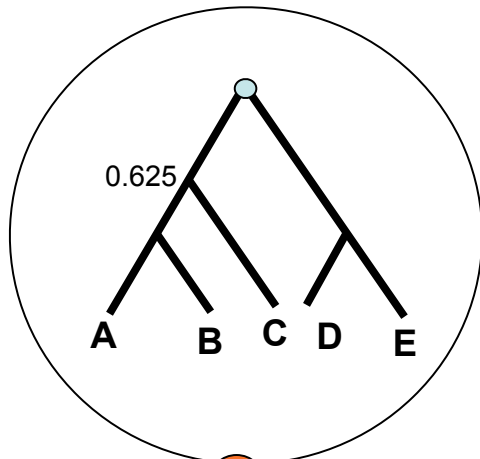
For each branchpoint in the parent tree,

For each tree, ask

Is there a branchpoint having the same subclade contents (i.e. same taxa, any order)

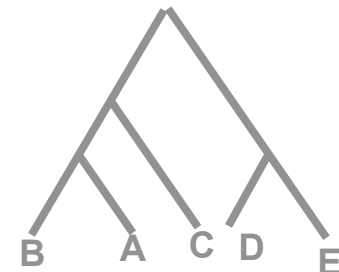
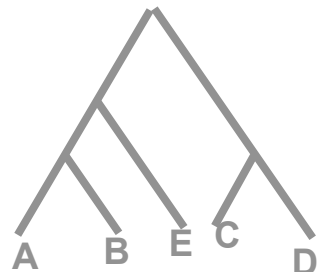
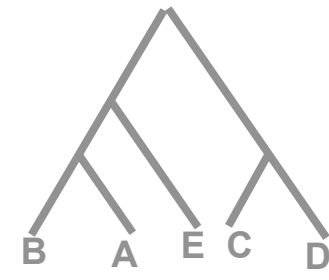
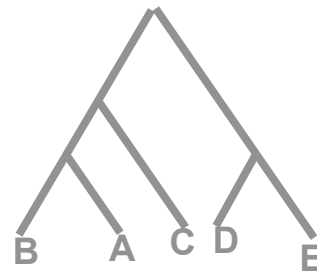
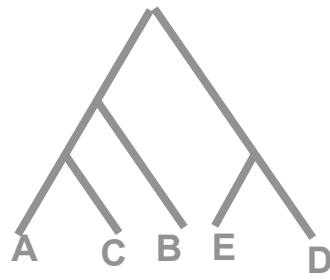
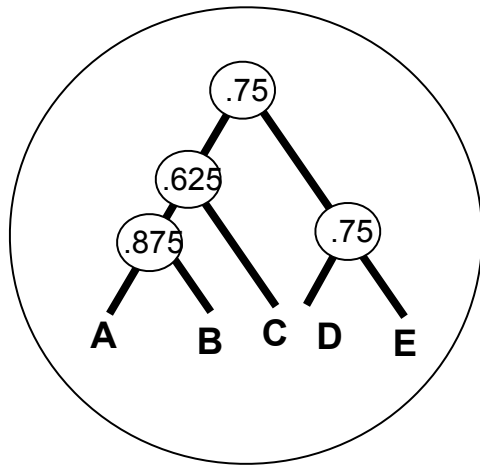
Bootstrap value = number of trees having the branchpoint / total trees.

Comparing branchpoints

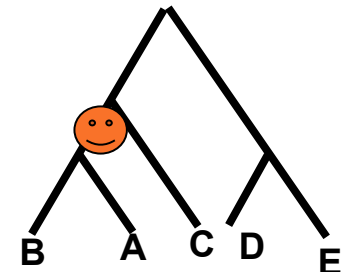
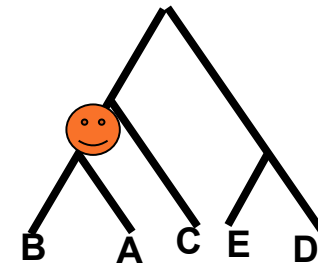
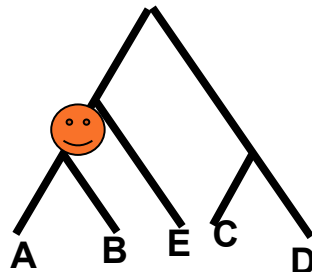
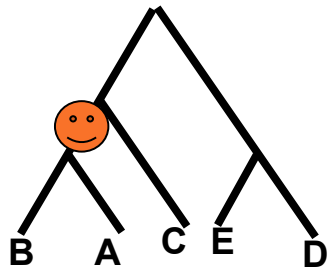
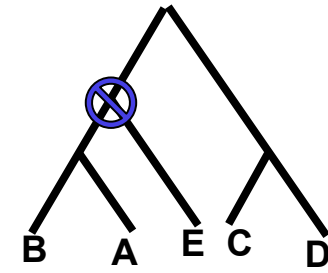
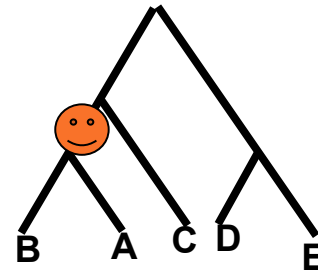
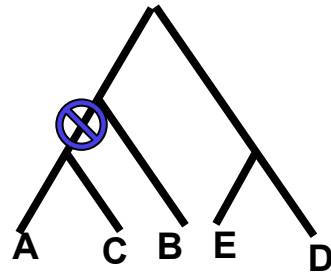
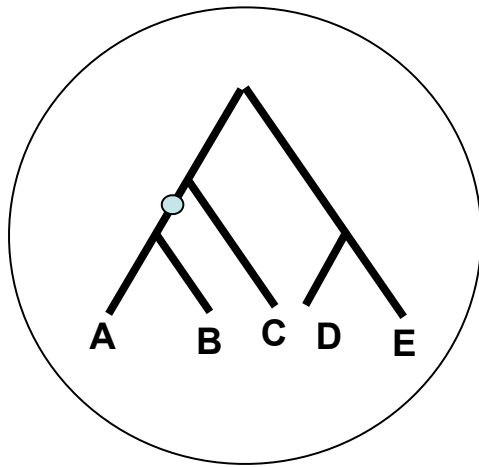


○ = $P((A,B,C),(D,E))=6/8$

Bootstrap values for this data



Compare lineages instead of branchpoints.



○ = $P(A, B) = 6/8$

For unrooted trees *or* rooted trees.

- Treat any lineage as the root.
- Ask how often the root branching is conserved.