Integrative Biology 200   "PRINCIPLES OF PHYLOGENETICS"   Spring 2016
University of California, Berkeley   Kipling Will

Tree selection, consensus, supertrees, and tree-to-tree comparisons

To choose or not to choose? In most cases an optimality criterion like parsimony or likelihood, will result in a set of equally most optimal trees and Bayesian methods are designed to produce a sampling of trees. In order to present your results in a publication a summary of the set of trees as one or a few figures is needed. Also, selecting one tree for subsequent analyses may be required by some applications. To select from a set of optimal trees, there are basically two stances.

1. *All trees of the optimal set are equal and so no single tree should be selected as preferable.*
    a. Alternative trees and tree islands may be discussed, but the strict consensus is presented as the "final word."
    b. If a procedure requires a single, fully resolved tree then one can be picked at random.
    c. Using a secondary optimality criterion to choose among the set of trees is not considered valid. Some would argue that if secondary optimality criteria can be justified, then it should be included in the initial search as a differential weighting or explicit model.

2. *Optimal trees represent a selected subset of all possible trees and implementing secondary optimality criteria to select from those trees is legitimate.*
    a. Subjectively preferred character state transformation
    b. a fit to external hypotheses or patterns for practical reasons
    c. explicit (numerical) secondary optimality criteria
        -successive Approximations Character Weighting (Farris 1969)
        - implied weights (Goloboff 1993)
        - average unit character consistency (AUCC) (Sang 1995)

sum ci/number of characters. This "packs" the most homoplasy in the fewest characters and thereby preserve the maximum number of initial hypotheses of homology.

Tree	Ci for characters 1 to 10										AUCC
T1	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	0.500
T2	1	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/3	0.533
T3	1	1	1	1/2	1/2	1/2	1/2	1/3	1/3	1/3	0.600
T4	1	1	1	1/2	1/2	1/2	1/2	1/2	1/2	1/5	0.620
T5	1	1	1	1	1/2	1/2	1/2	1/2	1/4	1/4	0.650
T6	1	1	1	1	1	1	1	1	1	1/11	0.909

Consensus: Representative summary of a set of source trees
The consensus efficiency can be calculated as CE =(log T- log C)/(log T- log S), T= number of trees for set of taxa; C number of trees allowed by the consensus tree; S number of source trees.

1. Strict consensus- Only monophyletic groups found in <u>all source trees</u> are found in the resultant tree. The tree excludes a subset of all possible trees and conversely includes a subset of possible trees, **whether or not** they are part of the source set. In some sense the most conservative consensus.

e.g. (A(B(CD))) + (A(C(BD))) = (A (BCD)) but this also implies or does not exclude (A(D(BC)))

-Length of a consensus tree is not comparable to the length on any one of the most optimal trees. Recall the best fit of a character to the worst possible tree, the bush.
-Characters optimized on a consensus tree may not reflect alternative character state changes. This is related to the consensus length issue above.

2. Semistrict (Bremer trees or combinable-components or loose) - Only monophyletic groups found in at least one source tree that are compatible (not in conflict) with all other source trees are found in the resultant tree, i.e. if a clade is never contradicted, but not always supported, then it is still included in the compromise tree.

3. Majority-rule - Shows groups that appear on pre-specified percentage of source trees, usually >50%. Used for summary of searches where plurality is important, e.g. Bayesian analyses.

	T1	T2	T3	T4	T5	T6	T7	TOT
AB								
CDE								
DE								
ABCDE								
XCDE								
XDE								
XC								
XAB								
XB								
XE								
XABCDE	1	1	1	1	1	1	1	7

4. Adams - Inconsistently placed taxa are moved to the first node that summarizes the possible topologies. Groups can appear in Adams consensus tree that are not found in any source tree. Adams trees have no biological or phylogenetic interpretation. They do point to "wildcard" taxa. Those taxa may be experimentally removed from the matrix and the resulting analysis compared to when they are included.

5. Greedy consensus. Groups ordered by frequency like in Majority-rule then added in to the consensus tree as long as they are compatible. How will ties in frequency change the results?

6. Matrix representation with parsimony (MRP). A recoding consensus method that can be used for trees with different sets of taxa. Both topology and frequency.

Supertrees
The objective is to construct trees that reflect the phylogenetic history of a large group (e.g. "the tree of life") by combining smaller source trees. It is a topology-based method and it does not use character data directly (and has been criticized for this). Importantly, source trees may have only partial overlap of OTUs.
Direct consensus methods similar to those above can be use, with an allowance for unequal sets

of OTUs, but more commonly indirect methods that have a tree coding step are used, e.g. MRP.

Since the input data for supertrees is phylogenies and topologies, it is possible to mix results from molecular and morphological analyses, or any tree source, to make up the supertree. This provides a general overview the major agreements among different studies and so may lend support to certain resolutions if all or most studies resolve a certain node in the same way.

However it has been criticized for potential bias because it enables the use of trees that include placement without analysis and republished trees that do not include any new data or analysis. Gatesy et al. (2002) give an example where MRP analysis done by Liu et al. (2001) resulted in a novel arrangement of mammals that apparently was an artifact of the method and data.

Tree-to-tree comparisons
Consensus methods tell us something about shared structure in trees. There are also many methods to compare two or more topologies to assess how different they are. These are frequently used in studies of co-evolution and gene tree/species tree relationships, for example.

a.Measures of difference: e.g., Quartet, Partition, Maximum Agreement Subtree, and Updown.
b. Transformations: e.g., Nearest Neighbor Interchange, Subtree prune, and regraft distance.
c. Reconcile tree analysis.

Farris, J.S., 1969. A Successive Approximations Approach to Character Weighting. Systematic Zoology. 18(4):374-385

Gatesy J, Matthee C, DeSalle R, Hayashi C. Resolution of a supertree/supermatrix paradox. Syst. Biol. 2002;51:652-664.

Goloboff, P. A. (1993). Estimating Character Weights During Tree Search. Cladistics 9: 83–91.

Liu, F.-G. R., M. M. Miyamoto, N. P. Freire, P. Q. Ong, M. R.Tennant, T. S. Young, and K. F. Gugel. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291:1786–1789.

Sang, T. (1995). New measure of distribution of homoplasy and reliability of parsimonious cladograms. Taxon. 44:77-82.
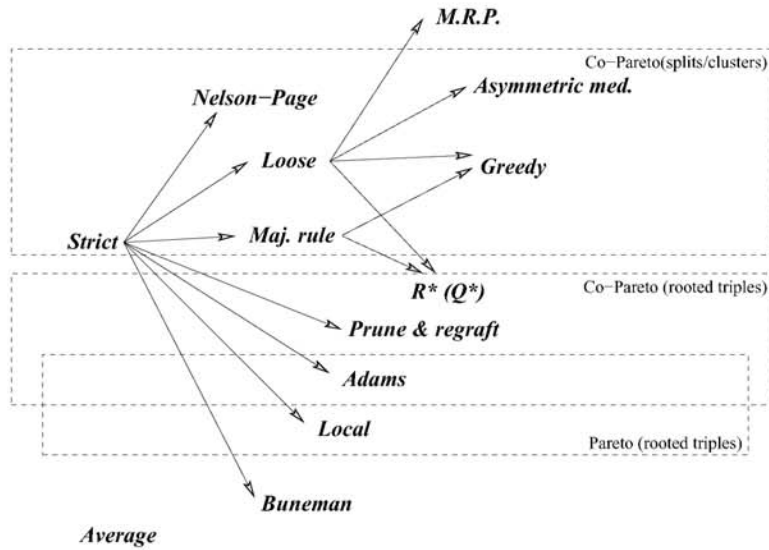
(k+1)/(si+k+1-mi)
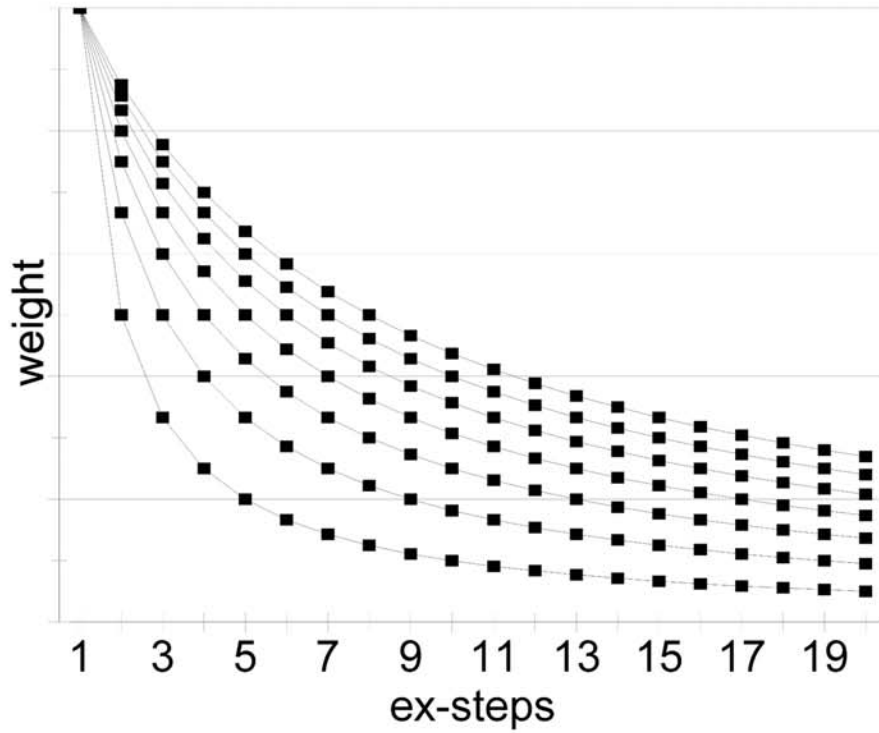
DAVID BRYANT

FIGURE 2. A classification of consensus methods. There is an arrow from one method to another if every split in the consensus tree produced by the first method is contained in every consensus tree produced by the second method.
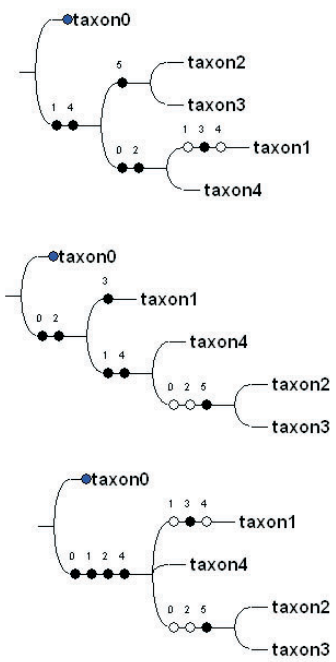
Fig. 1- two trees with length 8 and consensus with length 10. An example of why you should not consider length on the consensus tree.
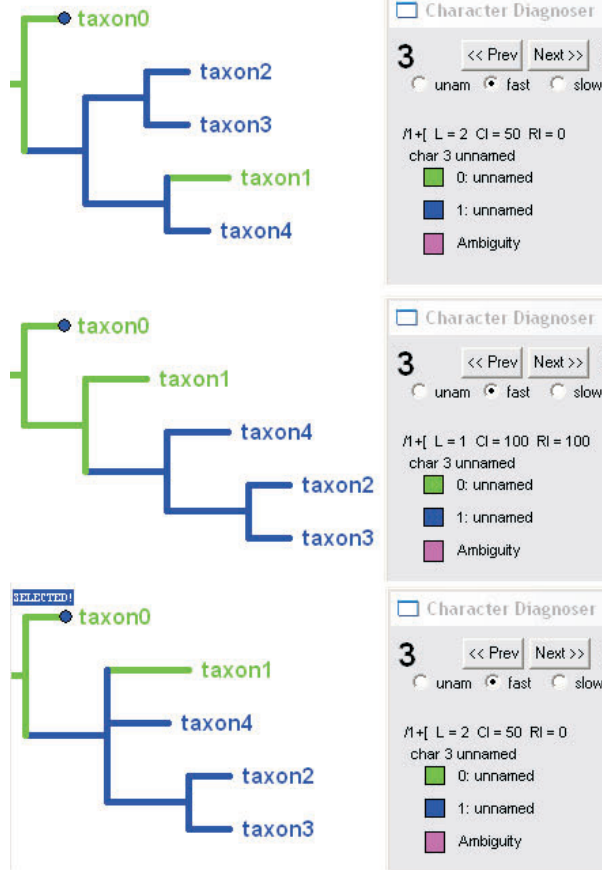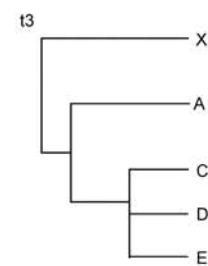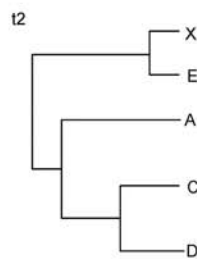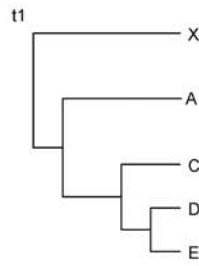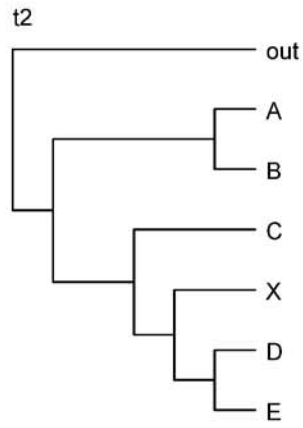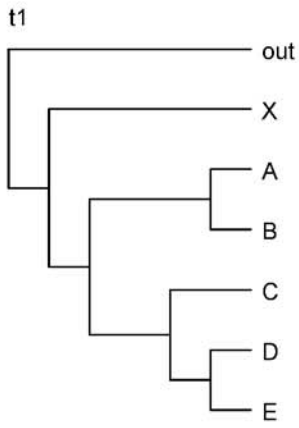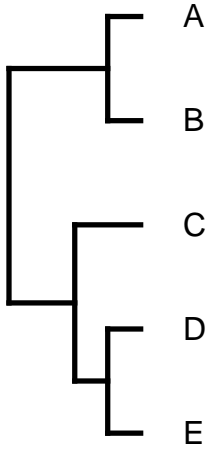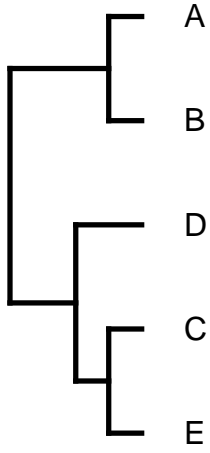

Fig. 2- An example of the problem of optimizing a character on the consensus tree.

A
B
C
D
E

A
B
D
C
E

A
B
E
C
D

A
B
C
D
E

B
A
C
D
E

E
A
B
C
D

A
D
E
B
C

(a) 1 2 3 4 5 outgroup 3 4 5 6 7 outgroup

B   C   D   F   E

A

(b)

Component

	A	B	C	D	E	F
outgroup	0	0	0	0	0	0
1	1	1	0	?	?	?
2	1	1	0	?	?	?
3	1	0	0	1	0	0
4	0	0	1	1	0	0
5	0	0	1	0	1	1
6	?	?	?	0	1	1
7	?	?	?	0	1	0

Taxon
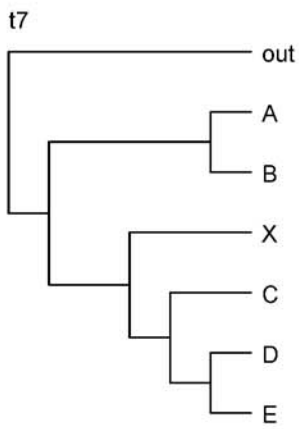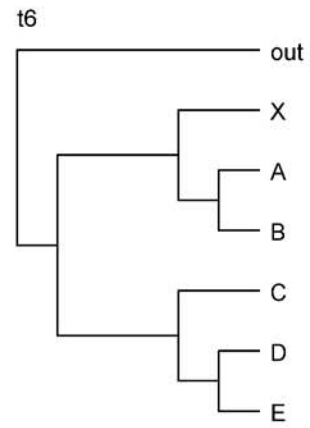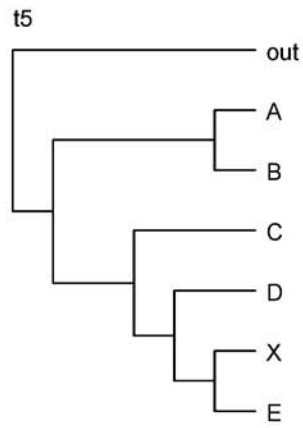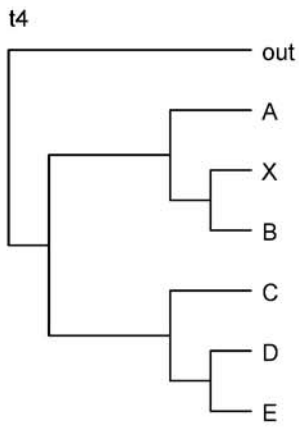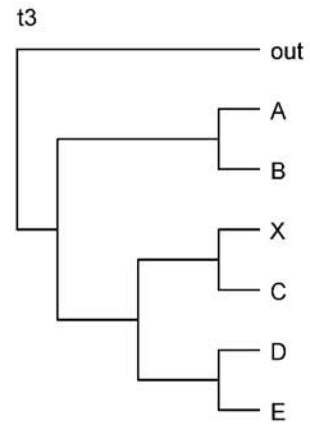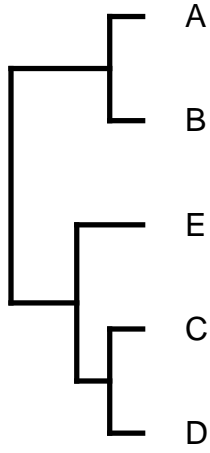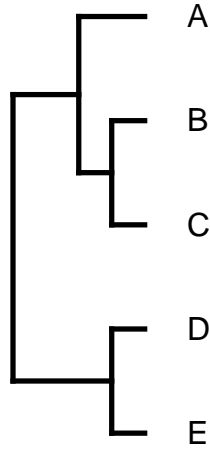
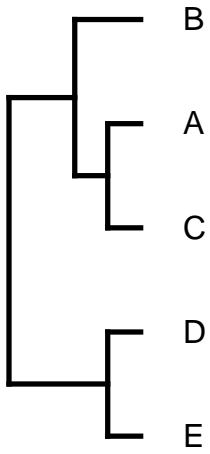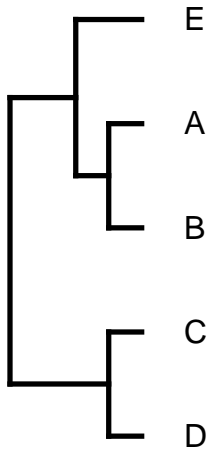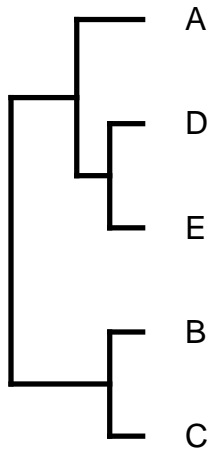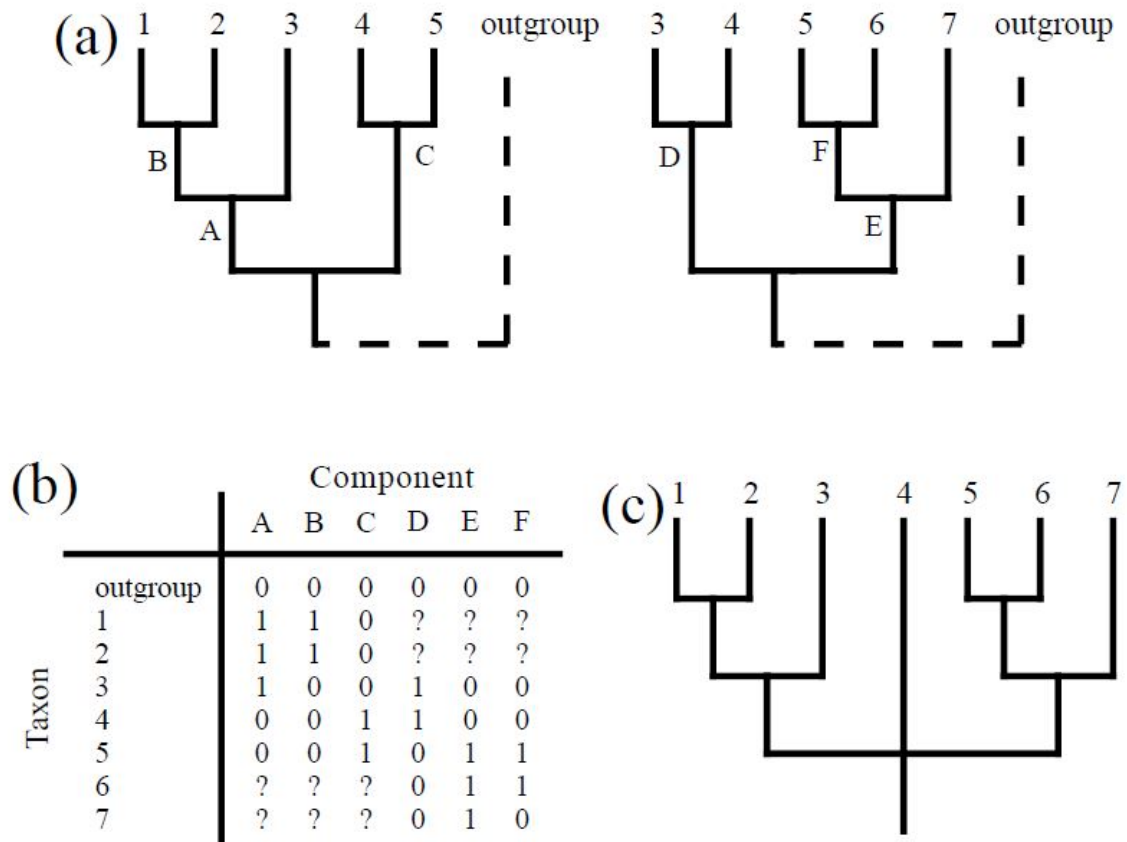(c) 1 2 3 4 5 6 7

Figure 2 Matrix representation with parsimony (MRP) supertree construction. (*a*) Two source trees for the taxa 1 to 7. A hypothetical outgroup has been added to each source tree. (*b*) The matrix representations of the source trees. (*c*) The MRP supertree.

Figure from Bininda-Emonds et al. 2002

Tree 1

Tree 2

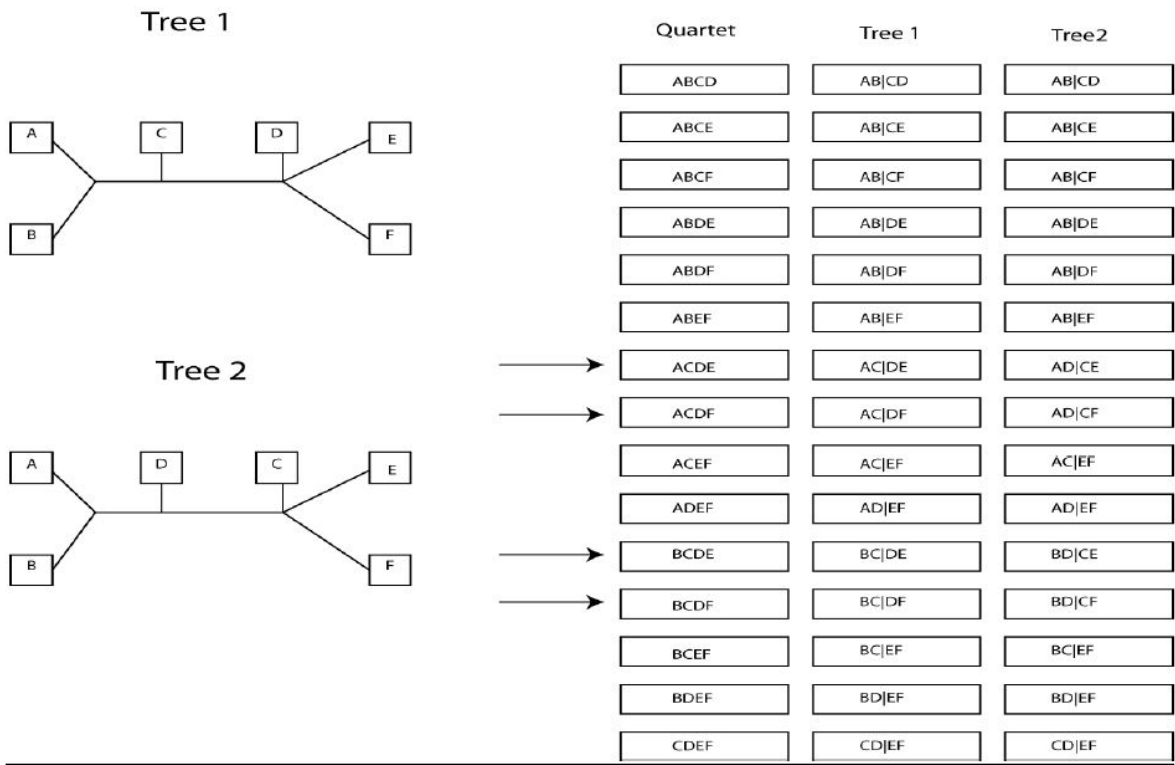| Quartet	Tree 1	Tree2
ABCD	AB|CD	AB|CD
ABCE	AB|CE	AB|CE
ABCF	AB|CF	AB|CF
ABDE	AB|DE	AB|DE
ABDF	AB|DF	AB|DF
ABEF	AB|EF	AB|EF
→ ACDE	AC|DE	AD|CE
→ ACDF	AC|DF	AD|CF
ACEF	AC|EF	AC|EF
ADEF	AD|EF	AD|EF
→ BCDE	BC|DE	BD|CE
→ BCDF	BC|DF	BD|CF
BCEF	BC|EF	BC|EF
BDEF	BD|EF	BD|EF
CDEF	CD|EF	CD|EF

Figure from Potiny (2010)