Integrative Biology 200A "PRINCIPLES OF PHYLOGENETICS" Spring 2012
University of California, Berkeley Kipling Will- 13 March

Tree selection, consensus, compromise

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 for some application. In terms of optimally, there are basically three stances one can take to do these things.

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 a random one should be used.
       c. Using a secondary optimality criterion to choose among the set of trees is not considered valid.
2. Optimal trees represent a selected subset of all possible trees and implementing secondary optimality criteria to select from those trees is legitimate.
       a. Internal evidence
              i. subjectively preferred character state transformation
              ii. explicit (numerical) secondary optimality criteria
•   Concentrating homoplasy and related methods
"the trees themselves tell us how reliable the characters are" (Goloboff 1993).

       <u>Successive Approximations Character Weighting</u> (Farris 1969)
- get starting MPTs
- use character fit to reweight (could be ci, ri or rci)
- search for MPTs with weights
- repeat until a stable set of trees is found.

IF this results in a subset of MPT from the original data those may be preferred. However, often this results in a different set of trees. It was not originally introduced to be a secondary optimality criterion to chose from among the MPT, but rather as a method for weighting characters. Consistent with this idea is <u>Implied Weights</u> of Goloboff (1993) and as implemented in TNT.

- weighting function is used to maximize weighted fit of characters to trees. $f_i = (k+1)/S_i + k + 1 - m_i$
k= constant (1..6); S_i = observed steps; m_i= minimum possible steps

e.g. For k=4 the cost of adding one step to a character with two extra steps is 54% of the cost to add a step to a "perfect" character.

-This kind of weighting function tends to push homoplasy into fewer characters and so the fittest tree(s) from the set of MPTs could be selected.

Related to these, but specifically for the purpose of choosing among MPTs is Sang's AUCC (1995).

sum ci/number of characters

Tree	ci										AUCC
A	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	0.500
B	1	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/3	0.533
C	1	1	1	1/2	1/2	1/2	1/2	1/3	1/3	1/3	0.600
D	1	1	1	1/2	1/2	1/2	1/2	1/2	1/2	1/5	0.620
E	1	1	1	1	1/2	1/2	1/2	1/2	1/4	1/4	0.650
F	1	1	1	1	1	1	1	1	1	1/11	0.909

- Pack the most homoplasy in the fewest characters and thereby preserve the maximum number of initial hypotheses of homology.
- Why this measure? Others abound.... optimal character compatibility index (OCCI) (Rodrigo 1992); boil-down (Sharkey 1989) etc.,etc.

        b. External evidence
                i. Correspondence to existing taxon hypotheses. Definitely the weakest and most problematic criterion.
                ii. Best fit to characteristics that are cannot be coded as characters.

3. If secondary optimality criteria can be justified, it should be included in the initial search.
                a. differential weighting or explicit model is used and you are back to number 1.

Consensus: Representative summary of a set 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. However, consider the bush.

e.g. (A(B(CD))) + (A(C(BD))) = (A (BCD)) but this also implies (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 (Fig.1).
-Characters optimized on a consensus tree may not reflect alternative character state changes. This is related to the consensus length issue above (Fig.2).

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.

NOTE: All trees below contain some resolution not supported in all source trees:

2. **Semistrict (Bremer trees or combinable-components)** - Only monophyletic groups found in at least one source trees and 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								
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 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.

(k+1)/(si+k+1-mi)

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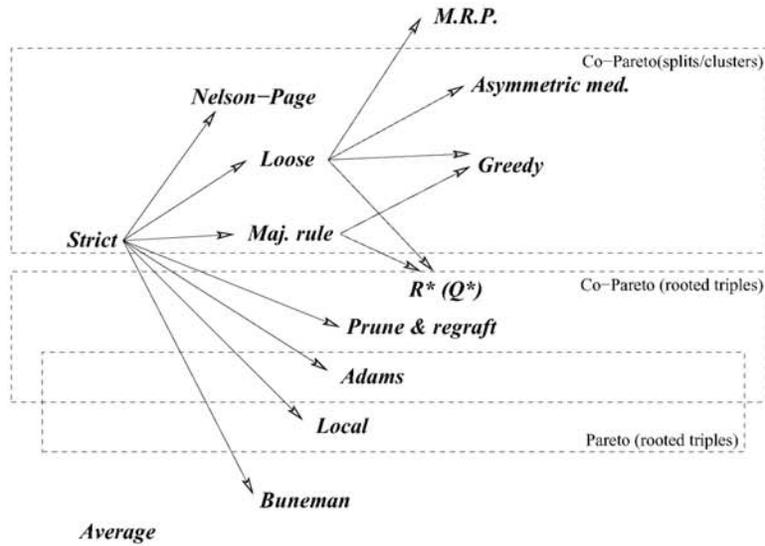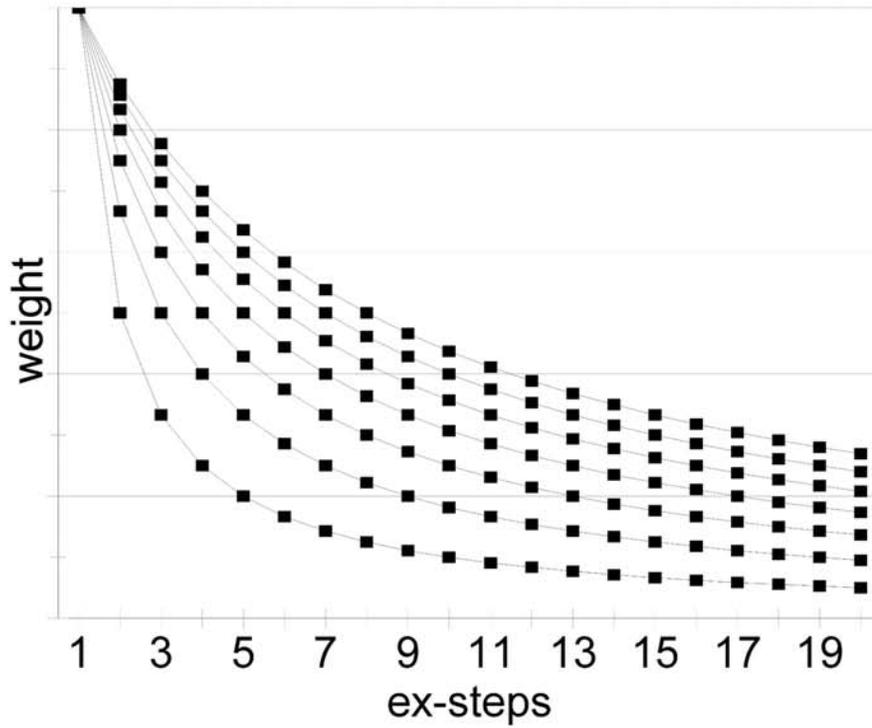
FIGURE 2. A classification of consensus methods. There is an
arrow from one method to another if every split in the consensus
tree produced by the first method is contained in every consensus
tree produced by the second method.
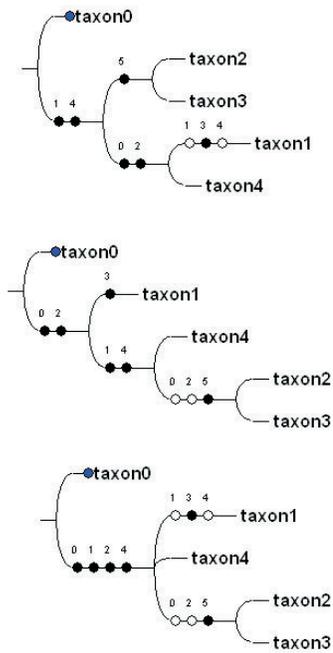
Fig. 1- two trees with length 8 and consensus with length 10. An example of why you should not consider length on the consensus tree.
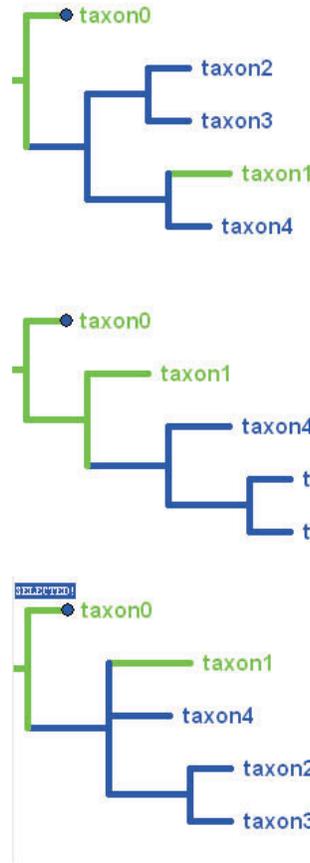

Fig. 2- An example of the problem of optimizing a character on the consensus tree.
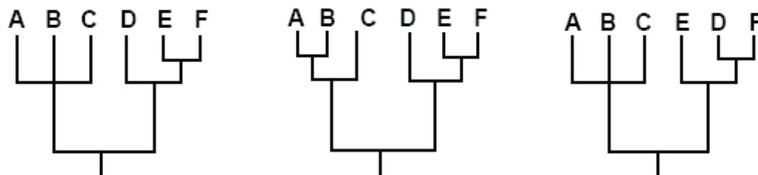
Fig. 3- Set of three trees with different results depending on methods used to make consensus tree.

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
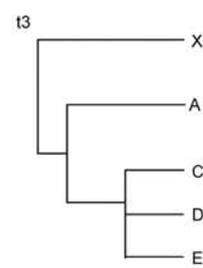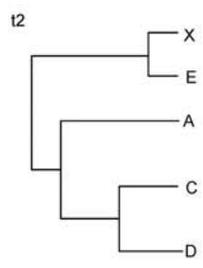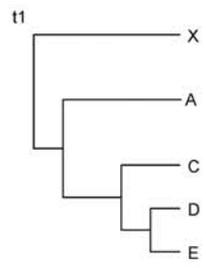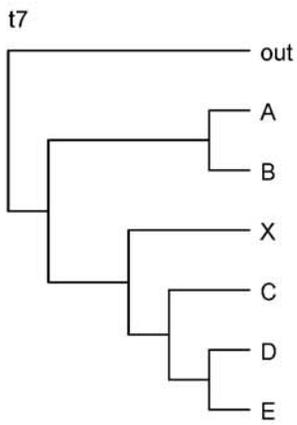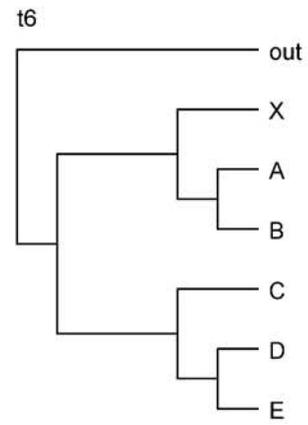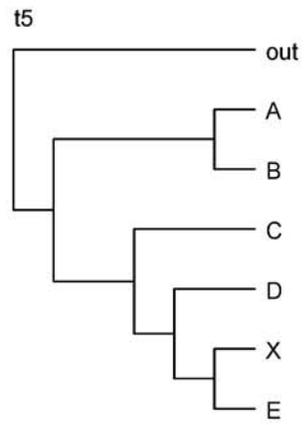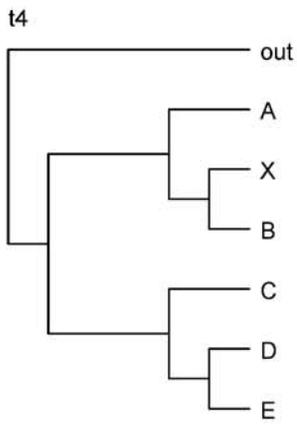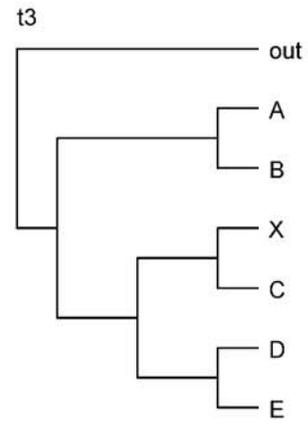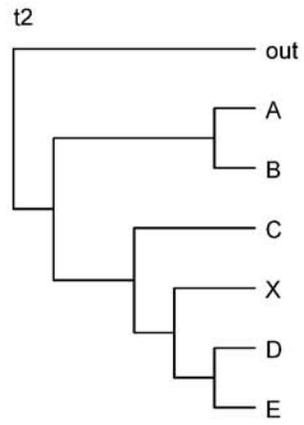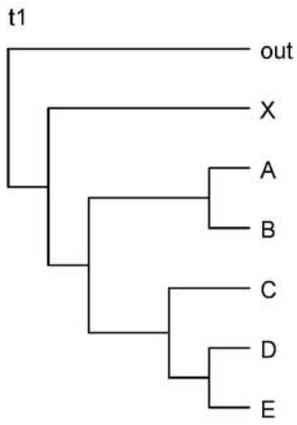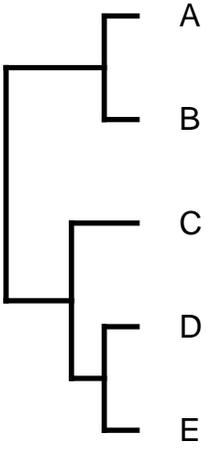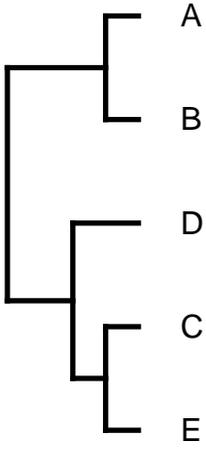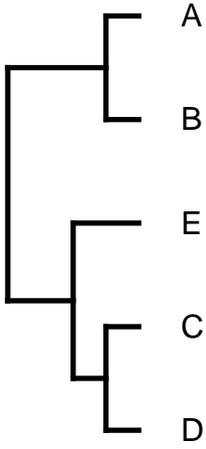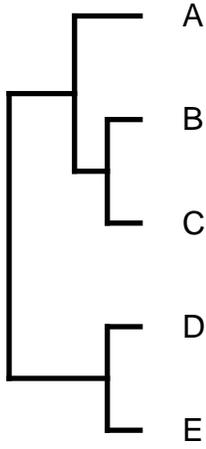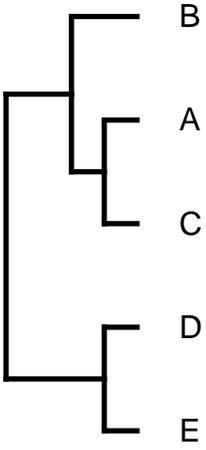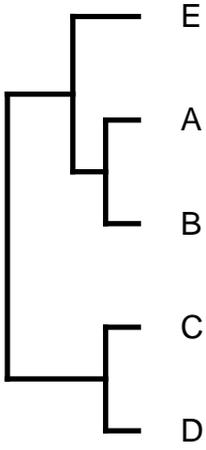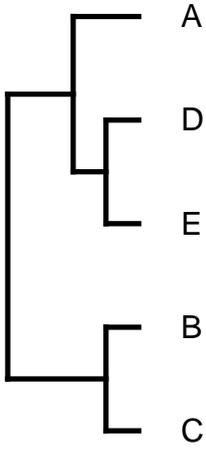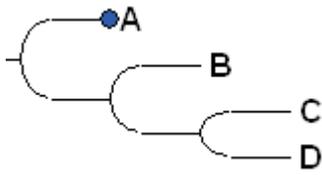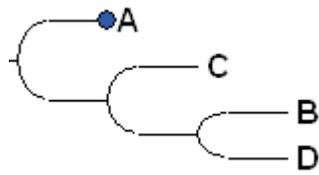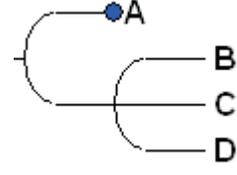
Tree 1

Tree 2

Consensus from Tree 1 and Tree 2
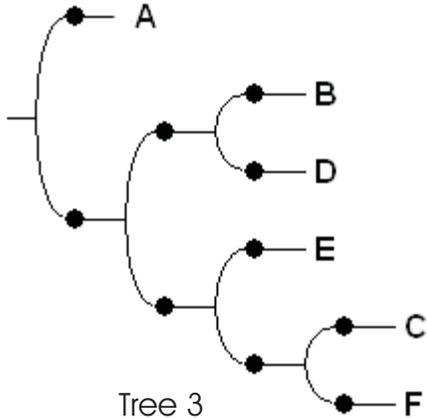
Tree 3

```
Matrices representing Tree 1,2,3
[M1,M2,M3]

        M1          M2          M3
A 1001000     1001000     10010000000
B 1100100     1110100     11101000000
C 1110010     1100010     11000100011
D 1110001     1110001     11100010000
E -------     -------     11000001010
F -------     -------     11000000111
```
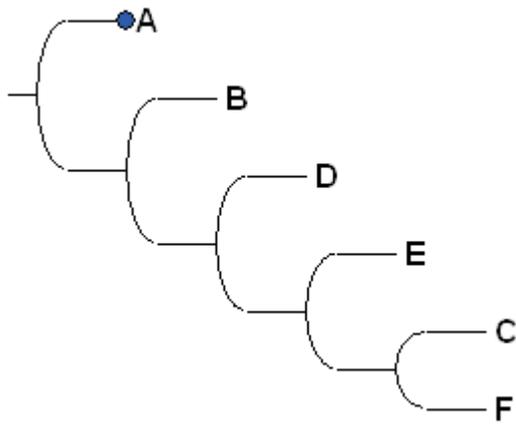


```
The result changes with two more topologies
like Tree1

        M1       M2       M3          M1a      M1b
A 1001000 1001000 10010000000 1001000 1001000
B 1100100 1110100 11101000000 1100100 1100100
C 1110010 1100010 11000100011 1110010 1110010
D 1110001 1110001 11100010000 1110001 1110001
E ------- ------- 11000001010 ------- -------
F ------- ------- 11000000111 ------- -------
```