Tree selection, consensus, compromise

Analyses will usually end up with more than one tree in the result set. You will face choices regarding how to present results, and how/when to choose a tree from among the set.

I. To choose or not to choose: There are basically three context dependent views on the way to proceed.

A. *If there are few trees examine them all,* also if there a few "kinds" of trees or "tree islands" as they are often referred to, examine one of each kind.

B. If you have far to many, large, complex trees to sift through then you may take the hard line. *All trees of the optimal set are equal. So no single tree should be selected.*

1. The strict consensus presented. You should do this for optimal sets of trees even if you select a tree or use other consensus methods.
2. If a procedure requires a single, fully resolved tree, a random one is used.
3. Using a secondary optimality criterion is not considered valid and (typically) equal weights were used. So if you think that secondary optimality criteria, weighting or a model could be used. Go back to the character analysis phase and explicitly include these.
4. As a process of reciprocal illumination, new data (characters and taxa) can be brought in and/or initial hypotheses revisited.

C. You have many trees but *optimal trees represent a select subset of all possible trees based on a conservative set of assumption and implementing a secondary optimality criterion(a) that selects from those trees is legitimate.*

1. Internal evidence
   a. subjectively preferred character state transformation
   b. explicit (numerical) secondary optimality criteria methods that concentrate homoplasy and related tree optimality methods. Use tree fit, weighting or homoplasy distribution to select a tree. (see below).
2. External evidence
   a. correspondence to existing taxon hypotheses (not popular here)
   b. best fit to characteristics that are cannot be coded as characters (but treated like this, these are never synapomorphies)

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

IF this results in a subset of MPT from the original data those are preferred. However, often this results in a different set of trees. Note that neither this nor PIWE below were originally introduced to be a secondary optimality criterion, but they have been used this way.

PIWE- Implied Weights (Goloboff 1993)
- weighting function is used to maximize weighted fit of characters to trees.
$fi=(k+1)/S_i+k+1-m_i$
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 and SAW tend to push homoplasy into fewer characters and so the fittest tree(s) from the set of MPTs *could* be selected.
- But if you think that maximizing retention of your initial homology statements is a reasonable way to select a tree, then choose by finding the tree(s) that pack the most homoplasy in the fewest characters. This preserves the maximum number of initial hypotheses of homology (character-wise) and you would choose the tree(s) with the highest AUCC.

AUCC- average unit character consistency, (Sang 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

- But….Why this measure? Others abound.... optimal character compatibility index (OCCI) (Rodrigo 1992); boil-down (Sharkey 1989). How different is a 0.500 vs. a 0.533 tree?

II. Consensus & Compromise: The representative summary of a set of source trees. Consensus trees can only be the most optimal tree when it is identical to one of the optimal source trees. Consensus trees have lost the information about what trees went into them, so reconstructing character evolution (mapping) and use of tree length on them should be avoided, or maybe done with extreme caution.

A. **Strict consensus**- A frequency based method. Only monophyletic groups found in all source trees 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 and this example...

e.g. (A(B(CD))) + (A(C(BD))) = (A (BCD)) but this also implies (A(D(BC)))

Note that the reason your consensus lacks resolution may be conflict, lack of character support or a combination of the two. Often authors mistakenly confuse these two.

All trees below contain some resolution not supported in all source trees. See figure from Bryant's (2003) paper:

B. **Semistrict-** A frequency based method. (aka, Bremer trees or combinable-components) - Only monophyletic groups found in at least **one** of the 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 this compromise tree.

C. **Majority-rule** – Again, a frequency based method. Shows groups that appear on pre-specified percentage of source trees, usually >50%. Used for summary of searches where plurality is important. Can result in a tree that contains two groups that are simultaneously found in **only one** of the source trees (minimum to make majority =0.5T+ 1).

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

D. **Greedy consensus**. Frequency based method. 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?

E. **Adams** – An intersection method. 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.

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

REF: Bryant, D. 2003. A classification of consensus methods for phylogenies. in Janowitz, M., Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (eds) BioConsensus, DIMACS. AMS. 163–184.