"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"

Integrative Biology 200B Spring 2009

University of California, Berkeley B.D. Mishler

April 14, 2009**. Comparing Cladograms; Supertrees**

-- There are many reasons why one would want to compare cladograms, falling into three basic categories:

-- *Within an analysis of one clade, with the same OTUs*; e.g., equally or nearly equally parsimonious (or likely) trees, trees resulting from different character partitions, models of evolution, or methods of analysis, and comparisons with trees from the literature.

-- *Within an analysis of one clade, with different OTUs*; trying to come up with a general tree for all OTUs, e.g. super trees, compartmentalization.

-- *Comparing analyses of different clades*, e.g., gene family evolution, migration between populations, vicariance biogeography, host/ parasite relationships, symbiosis, community evolution, or any long-term ecological association

-- Methodology for comparing cladograms:

(1) consensus techniques (strict, semi-strict, majority rule, Adams) -- for finding shared signal among trees.

Strict consensus: 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, e.g. (A(B(CD))) + (A(C(BD))) = (A (BCD)) but this also implies (A(D(BC))). In some sense the most conservative consensus. However, consider the bush.

Semistrict consensus: 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. E.g. (A(B(CD))) + (A (BCD)) = (A(B(CD)))

Majority-rule consensus: Shows groups that appear in more than a pre-specified percentage of source trees, usually >50%. Not recommended for summary of equally-optimal trees resulting from a search.

Adams Consensus: Inconsistently placed taxa are moved down to the first node that summarizes the possible topologies. N.B., groups can appear in Adams consensus that are not found in **any** source tree. Adams trees have no biological or phylogenetic interpretation, but 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.

(2) tree-to-tree distance metrics. There are two types of approaches. One counts the number of steps needed to transform one tree into another (e.g., NNI interchange metric, partition metrics, agreement subtrees). The second represents two trees as sets of simpler structures and then measures similarity between these (e.g., quartet measures)

Transforming one tree into another

A good example of a measure defined in terms of transforming one tree into another is the nearest nearest neighbor interchange (NNI) metric (e.g., Waterman and Smith, 1978) which measures the minimum number of NNIs required to change T_1 into T_2. In the example below, one NNI is required to convert T_1 into T_2, so $d_{NNI}(T_1, T_2) = 1$.

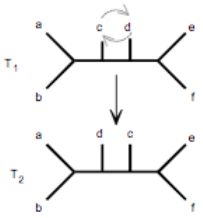
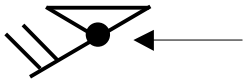


Figure 5.1 Transforming T_1 into T_2 by a single nearest neighbor interchange of leaves c nd d

from the Component User's Guide, by Rod Page (http://taxonomy.zoology.gla.ac.uk/rod/cplite/title.pdf)

(3) component analysis (more next week in biogeography) -- finding individual statements of relationship that are shared among trees, basically a node relating some taxa to the exclusion of others.



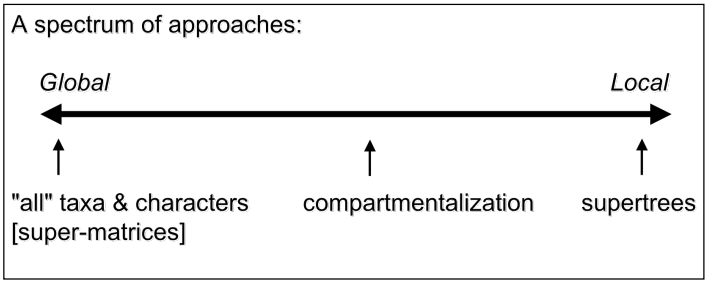
(4) maximum likelihood approaches (parametric bootstrapping) -- comparing alternative trees or alternative models of evolution for your data.

(5) representing the grouping information in separate trees as characters in a matrix (e.g., using Brooks parsimony, also called "matrix representation parsimony"). This might be used when comparing hosts and parasites, or phylogenies of different taxa that all live in the same areas of endemicity. Brooks' approach is be gone over on the last page, and the chalkboard.

(6) *supertrees* are one of the frequent applications of tree comparisons, in this case attempting to combine different trees of the same larger clade that were developed from different sets of OTUs. In the simplest case, detailed phylogenies of individual genera are stitched together using a backbone phylogeny of a family that might have one representative of each genus. For a more analytical approach, Brooks parsimony can be used (branches in the separate trees are represented in a data matrix for analysis).

(7) *compartmentalization* is related, but differs in having actual analyses at each step. Procedures in compartmentalization:

i. global analysis, determine best supported clades (= compartments)
ii. local analyses <u>within</u> compartments, often with augmented data sets
iii. return to global analyses, either:
    (a) with compartments constrained to local topology (for smaller data sets); or
    (b) with compartments represented by a single HTU -- the inferred archetype



Brooks Parsimony

(see Brooks & McLennan, 1991; Brooks 1981, Syst. Zool. 30:229; Wiley 1988, Syst. Zool. 37:271; and see Kluge 1988, Syst. Zool. 37:315 for some suggested modifications)

Steps:

1. Cladogram of parasite group

2. Cladogram of host group

3. Cladogram of parasite group is taken as a completely polarized, multistate transformation series -- recoded by additive binary coding

4. make new data matrix with hosts as OTU's and parasite clades as characters

5. construct new host cladogram from this matrix

6. compare this new host cladogram (derived from cladogram of the parasite group) with the original host cladogram (derived from host characters) and with host cladograms based on other parasite groups, if possible. Congruence is taken as evidence of common cause (shared history); incongruence (homoplasy) is taken as due to separate causes (e.g., host-switching or extinction)

-- Problem resolution:

1. when more than one parasite occurs in given host, codes are combined (e.g., host E in example combines codes of 10 & 5)

2. when all members of parasite clade are missing from a host taxon, host coded with "?"

3. when one parasite occurs in more than one host, codes are combined (e.g., if parasite species 10 occurred in hosts A,B,& E) -- note that this is controversial -- some have suggested downweighting such species, even eliminating them from the analysis entirely.