Data/Hypothesis Exploration and Support Measures

I. Overview.

The simplest means of establishing confidence in our phylogenetic hypotheses, though somewhat subjective, is to show character state changes on the cladogram. Groups supported by a large number of less homoplastic and more complex character state changes are thought to be well supported. When more of our initial statements of homology survive the test and are compatible we have increased confidence in the hypothesis. This may only be well suited to morphological data, however, as it is more difficult to apply DNA sequence data given the simplicity of the states.

Even a general or specific "fit" to data external to the analysis (e.g. biogeographic patterns, ecological aspects) builds confidence. However, this is narrative and subjective, making it hard to evaluate especially if you are not actively working within the taxon. It is necessary to express some sort of confidence or make a statement of reliability in order to give *others* a sense of how well your data fit your hypothesis and to what degree the critical evidence refutes competing hypotheses even if we are confident in the result.

Many agree that the best tests in phylogenetics involve empirical tests that examine all critical evidence at hand. For phylogenetic hypotheses (tree, branching pattern, branch lengths, character state distribution), this involves the addition of more characters and taxa. This is not always reasonable or feasible. When do we have enough anyway? This is an issue of philosophical and/or statistical confidence.

Typically "exploration" methods seek some sort of statistical reliability as an estimate of accuracy or measure giving an impression of how bold or conservative we should be in regard to conclusions based on the phylogenetic pattern. The fact that nearby, sub-optimal solutions exist is not enough to cause us to move from one hypothesis to another.

Although there is a general notion that we are identifying well supported clades, exploration methods and support measures are really just as, or more important for pointing to poorly supported parts of the tree. Poorly supported groups suggest where future efforts need to be applied.

Perhaps one of the most controversial aspects of common support measures is that they, like most statistical methods, require assumptions about the universe from which the sample is drawn. Generally this is assumed to be a random sample of the universe of possible independent entities, i.e. they are *independent and identically distributed* (i.i.d). Whether this holds for a phylogenetic character has been debated.

II. Sensitivity and Resampling Analyses: Various, usually heuristic methods explore how robust the hypothesis is if the underlying assumptions are wrong or data and OTUs are altered.

A. Methodological concordance: Multiple methods of phylogenetic analysis are used and the clades found in common are presumed well supported. Controversy over methods and assumption can be avoided by a pluralistic approach that leads to reasonable results. Accurate methods will converge on the "truth" and a lack of agreement between methods indicates that none are recovering the true tree (Kim 1993). *Note that I do not agree with any of what was just stated, but apparently some people do.

What doing this really tells us is which clades are not affected by the assumptions that differ between the methods that were used for analysis of the data. Since various methods address the problem from very different statistical and philosophical views, the fact that they converge may say something about the data or the methods (possibly tell us if long branch attraction is a problem) but may have little to

do with discovery of "correct" groups. There is no clear connection between convergence of methods and "accuracy." All accurate methods should converge on the truth; however, convergence of methods does not necessarily mean they are accurate. Maybe they are all wrong.

B. Assumption sensitivity analyses: Assumptions (parameters) are varied in multiple analyses and the results compared in some way. This is done to look for sensitivity to variation in model assumption (e.g., different weights assigned to transitions/ transversions changes topology, alignments, etc.). This has been used as an optimality criterion for deciding if a group should be accepted or rejected. Groups sensitive to variation are rejected. Also used as a means to select a set of alignment parameters.

It shows which groups remain under a set of "reasonable" parameters and support is drawn from a variety of synapomorphy classes. Not truly a test of monophyly or support. Monophyly is tested in the corroboration of empirical evidence in light of some set of "valid" assumption. It doesn't really test the support offered by the data. Almost any topology can be supported under some set of parameters. It can't distinguish levels or different kinds of support and a group well supported under "mutation-parsimony" may be lacking under "in/del- parsimony"

((a,b) [10,10,10] ((c,d) [1,1,1](e,f) [30,0,0])) or (a,b) [10,10,10] vs. (a, e) [30,0,0]

C. Bremer Support / Decay analyses or "index" Record the number of extra steps required to lose a clade that is found in the most parsimonious tree. Any clade not found in the strict consensus of all MPTs has a Bremer support value of 0. Any clade not found in the strict consensus of all trees one step longer than the MTPs has a Bremer support value of 1,2,3... until a shortest tree that does not contain any clade is found. In reality this typically includes far too many trees to have an exhaustive set and a Bremer support value is an *estimate* based on *heuristic* searches of suboptimal tree space. As such it is important to have each of the searches in the decay analysis be as, or near as rigorous as the primary tree search.

This gives a measure of decisiveness and indicates ambiguously supported nodes directly from the data. It provides an estimate of the degree to which the optimal solution is preferred to alternatives. As a heuristic, it points to poorly supported groups that may have few synapomorphies or nearby alternatives may be supported by conflicting characters. However, it does not discriminate between different types of support and does not have a clear statistical interpretation.

Matrix w/100 characters and MPTs of 200 steps, each character optimizes for 2 steps. Trees 2 steps longer (202 steps) could come by increasing one character to 4 steps [99*2 +1*4= 202] OR reducing 49 characters to 1 step and increasing 51 to 3 steps [49*1 + 51*3= 202].

D. Bootstrap/Jackknife: Resample with replacement (bootstrap) or without replacement (jackknife) from the matrix. Essentially Bootstrap differentially weights some characters to build a matrix of equal size. Jackknife reduced some characters to weight of zero. For either method, the resulting matrix is used to build a set of trees. This is repeated many times to build a cloud of trees. A majority rule consensus tree for groups found in >50% of the trees is used to show well-supported groups.

In phylogenetics these methods are used to assess uncertainty in the proposed phylogeny and are usually applied to characters but also have been used to resample taxa.

The issue of independence of sampled elements is debatable, but generally in statistics these are only used for random samples that are independent. For taxa, most people agree that because they are more or less phylogenetically related they do not represent i.i.d. samples. However, some (e.g. Felsenstein) maintains that characters are less likely to be non-independent than taxa (an assumption made most of the time) or this can be corrected for. Some maintain that a sample of characters in a matrix is not drawn from an i.i.d. of all possible characters and this

invalidates the method for phylogenetic characters. Other say that the sample need only be drawn independently from "some" universe of characters. But if the Bootstrap tree is different than the sample universe of empirical data that we have, it must be a poor estimator. Characters that are not parsimony informative are potentially problematic. Bootstrap has been shown to be positively correlated to number of informative characters (parsimony); negatively correlated to number of taxa in analysis, number of taxa in a clade and tree asymmetry (Siddall 2002). Also there is autocorrelation of nested clades (e.g. clade (D,E,) and its supporting characters are not independent of (C(D,E)).)

Re-sampling biased data would only lead to an assessment of the accuracy of the bias. At best, as a heuristic it points to poorly supported groups that may have few synapomorphies or may be supported by conflicting characters.

III. Permutation tests. Tests for hierarchical structure and phylogenetic "signal" to reject the null of no structure or reveal conflicting structure.

A. PTP, etc. Character state data is randomly and independently reshuffled among taxa, optimal trees are found for each permutation and compared to establish confidence limits, e.g., 95%. Either tree length (permutation tail probability- PTP) may be used or the clades (topological dependent permutation tail probability- T-PTP) may be compared. This places confidence limits on the clades relative to Type-1 error (errors resulting from wrongly rejecting the null hypothesis = no structure.). If the optimal score for the original data is far out in the distribution tail then significant, non-random structure is present in the data. However, PTP can show significant support for a group that has none in the original data. A single resolved node (either an internal polytomy or a pair of very close species) may give a significant result for an otherwise unstructured data set. Similarly the T-PTP has a null hypothesis that there is no structure in the matrix anywhere, so it is likely to reject the null too easily.

B. Skewness: Look at the number of changes on all possible topologies (actually a random sample). If there are a few trees of much lower score they will negatively skew the distribution. Strongly skewed distribution suggests the "strength" of the phylogenetic signal or decisiveness in the matrix. Hard to tell what this really tells us. A number of published examples show it may fail to reflect phylogenetic structure and it is influenced by the central mass of the distribution more than the tail, influenced by character state distribution and requires arbitrarily resolved polytomies. Nevertheless, it is still being used in publications. Perhaps, because for no other reason than it is available in the Paup menu.

C. ILD (incongruence length difference): Compare the length of the most parsimonious trees for two or more data matrices or partitions to their length in the combined analysis and/or to randomly sampled partitions of equal size.

$$ILD = L_{AB}-(L_A + L_B)/L_{AB}$$

This has been suggested as a test to determine is some data partition should be excluded or reweighted rather than direct and equal combination. Has been used to select which alignments parameters to use or what model should be preferred for a partition (an interesting paper on this is Aagesen et al. 2005).

Significant incongruence suggests that partitions may have a different evolutionary history. Some people would not combine data that showed significant incongruence. However, without evidence that there is some process that would cause the incongruence, down-weighting or eliminating character data simply because it is incongruent is really not well justified. Examining RI and CI (see below) of partitions

would probably be as informative and would allow for all data partitions to be examined in light of all critical evidence in a combined analysis.

IV. Basic Descriptive Indices:

Consistency Index (CI & ci)

Measure of how data fits the tree topology. Give the amount of homoplasy in a character or matrix for a give tree.

ci = m/s -----where m = minimum number of steps in a character (number of states -1)
s = steps actually realized on a given tree

e.g. binary character m=1 actually has 1step on the tree then ci=1.0 if it has 2steps on the tree then ci=0.5

This index falls between 0 and 1.0 but is usually reported as scaled between 0-100

Ensemble CI (for the whole matrix) is the sum of all m/ total length of the tree (CI=M/S). In general, a high CI indicates that the data matrix "fits" the tree well (i.e., contains little homoplasy for the particular tree topology), whereas a low CI does not.

Characters with the same ci may not be contributing to the tree topology equally (e.g., autapomorphies ci=1.0), so CI may be an overestimate if these are included. CI is NOT comparable between different sets of taxa as more taxa decreases CI.

Retention Index (RI & ri)

Measure grouping information in the data.

ri = (g – s)/(g – m) ------where g= minimum steps on the worst tree
(=bush) Ensemble RI (for the whole matrix) like CI is based on sums
RI=(G-S)/(G-M)

These problems for CI noted above may be overcome by excluding autapomorphies OR calculating a Rescaled Consistency Index. RC = RI*CI

This removes the impact of any characters that do not contribute to the "fit" of the data to the tree (e.g., autapomorphies ci=1.0 and ri=0.0)

WHAT THESE TELLS US: These describe aspects of the tree and matrix or partitions of the matrix (e.g. 3rd position might have a lower CI and/or RI than 1st) or a particular sequence may contribute more to the resolution that another.

Aagesen, L. Petersen, g. and O. Seberg. 2005. Sequence length variation, indel costs, and congruence in sensitivity analysis. Cladistics. 21:15-30.

Kim, J., 1993. Improving the accuracy of phylogenetic estimation by combining different methods. Syst. Biol. 42, 331–340.

Siddall, M. E. 2002. Measures of support, in R. DeSalle, G.Giribet & W. Wheeler (eds.), Techniques in MolecularSystematics and Evolution, pp. 80-101. Switzerland, Bir-khaüser, Verlag.