Phylogenetic Congruence

I. What to do with multiple data sets

A. "Total Evidence" - Combine always
   1. any data partition is arbitrary
   2. across data sets signal [synapomorphy] will be amplified over randomly distributed noise [homoplasy]

B. "Consensus" - analyze separately, and assess level of congruence
   1. views separate analyses as independent tests
   2. will result in a more conservative [less resolved] estimate of phylogeny

C. "Conditional Combination" - combine data sets except when significant heterogeneity exists among them
   1. accepts the desirability of total evidence
   2. BUT, some conflict is NOT random e.g. interspecific gene flow, gene conversion

II. Topological Congruence

A. Tree interconversion

B. Subtree similarity

C. Shared clades identified by consensus methods

D. Limitations
   1. Do not measure the statistical significance of incongruence
   2. Do not identify the source (i.e. characters) responsible for incongruence

III. Character Congruence

A. $I_{MF}$ = Michevich-Farris Incongruence (1981)
"the proportion of extra homoplasy required to explain an alternative topology"

0 = shortest trees from each data set are identical

1 = both data sets have no homoplasy, and the topologies are unique

B. Partition Homogeneity Significance Test

1. null hypothesis = homogeneity in the distribution of phylogenetic information

   a. compute length tree A + tree B
   b. create random partition
   c. compute length of tree A' + tree B'
   d. longer length = smaller $I_{MF}$
   e. if >95% of random partitions give a greater length, than trees are significantly incongruent

C. Significantly Less-Parsimonious Test
   (Templeton 1983)

1. Each character on alternative tree will require the same, fewer, or more steps

2. Null hypothesis = the increase in the # of steps shown by some characters on the alternative topology is not significantly greater than the decrease in steps shown by other characters

3. Uses Wilcoxon signed-rank test

IV. Kishino-Hasegawa (1989)

A. The sum of a collection of random variables is approx. normally distributed
   [Central Limit Theorem]

B. The elements of the sum are the ln-likelihoods of each nucleotide site [for a particular tree]

C. Variation among ln-likelihoods used to estimate variance of their sums

D. Resulting variance used to test statistical significance of two trees
V. Applications

A. Comparing data sets
e.g. molecules vs. morphology, genes from different genomes, multiple genes

B. Comparing alternative topologies [Templeton & K-H]
e.g. less parsimonious trees, trees from other analyses, trees drawn from existing classifications