

Consensus Trees

- * consensus trees reconcile clades from different trees
- * consensus is a conservative estimate of phylogeny that emphasizes points of agreement
- * philosophy: agreement among data sets is more important than agreement within data sets
- * a position of safety
 - defensible and pragmatic starting point... especially if you are proposing a new classification or testing a hypothesis

Consensus Trees

(1) Different data sets

same taxa; different character systems

e.g., larval and adult data for insects

e.g., molecular sequence data versus morphology

- prevent molecular characters from swamping out morphological data

e.g., must use consensus methods for some data sets;

- distance plus discrete character data

(2) Comparing results from different algorithms

same taxa, same data; different algorithms

e.g., distance vs parsimony or likelihood trees

- one scenario here is when you have some long branch problems and algorithms deal with them differently

Consensus Trees

- (3) Choosing among trees of equal stature
same taxa, same data, same algorithm, different trees
e.g., have set of equally MPTs, but need a summary
solution
e.g., need to summarizing set of bootstrap replicates of
your data

Note: even when topologies are exactly the same tree can
differ in

- * how character are plotted (reconstructed) on the trees
- * how branch lengths are fitted

Consensus Methods

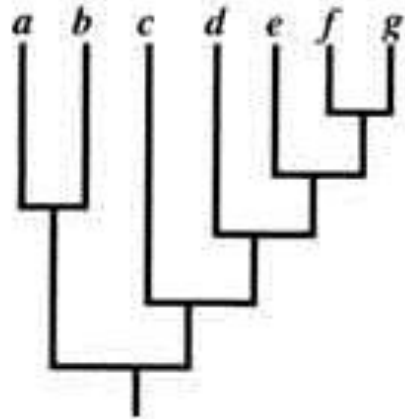
Label all components: each distinct component (clade) is given a unique number. Different algorithms/methods work with these numbers (have different rules)

***Strict Consensus (Nelson 1979, Sokal & Rohlf 1981):** only those components (clades) shared by all trees are considered; components must be exactly replicated among all trees. Most restrictive approach.

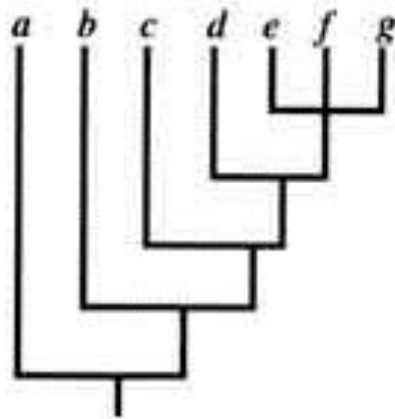
***Consensus n-Trees (Margush & McMorris 1981):** accepts all nodes/resolutions that are present in n% or more of the trees. Usually n=50 and referred to as *majority rule consensus*.

Adam's Consensus (Adams 1972, McMorris et al. 1982): pulls down components to the first node to which there will be no conflict. Most unrestrictive approach. Preserves structure.

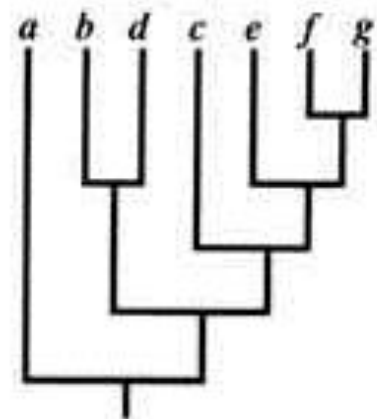
* You are only responsible for the first two



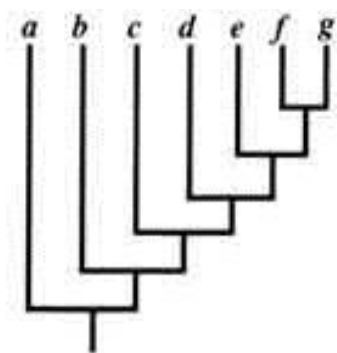
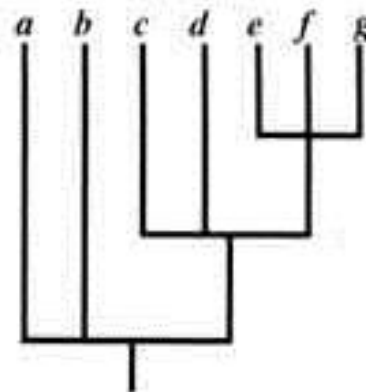
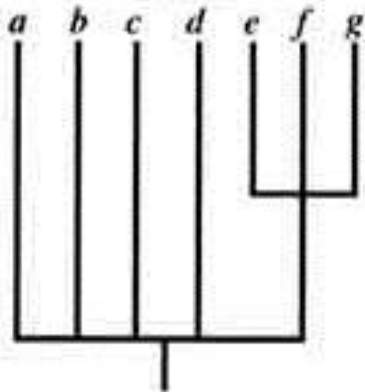
strict



Adam's

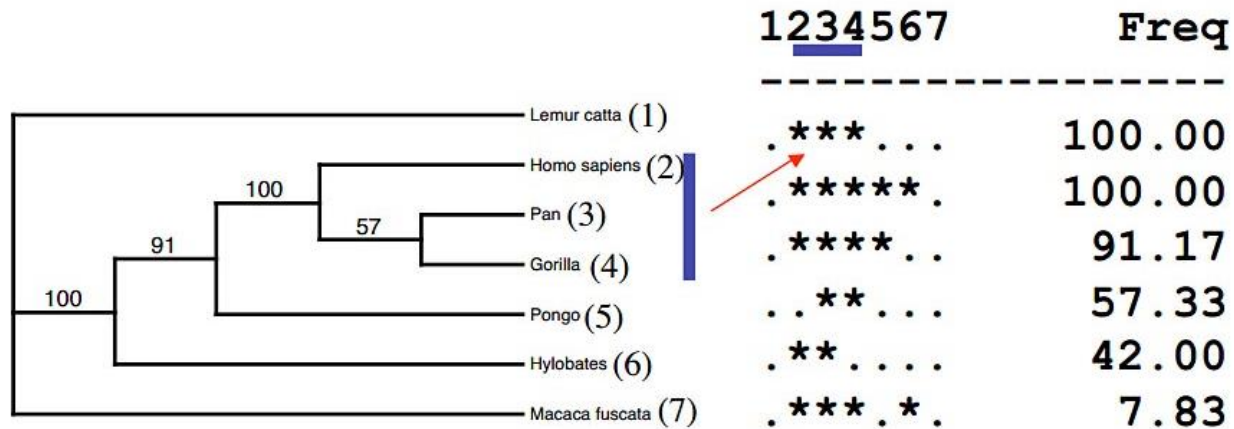


majority rule



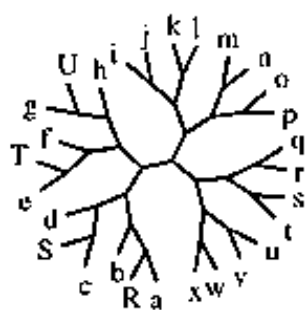
From Quicke 1993. Principles and Techniques of Contemporary Taxonomy

Majority-rule consensus

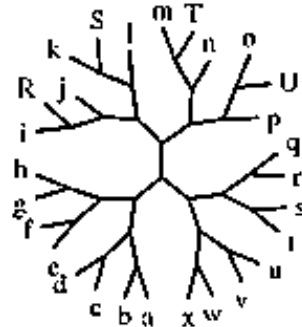


PAUP output

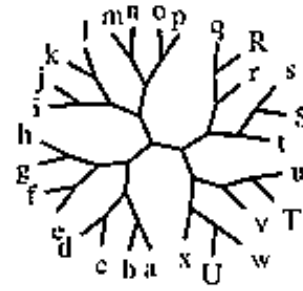
Contrived example with rogue taxa



(a) Tree 1



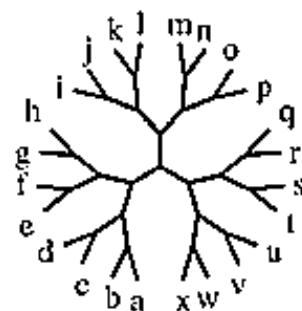
(b) Tree 2



(c) Tree 3



(d) $C_{m-1}(T)$



(e) $C_{m-1}(T|\{a, \dots, x\})$

A simple, yet starkly contrasting, example for which the strict consensus [of a, b, and c] returns a star tree, but for which our algorithm correctly identifies the rogue taxa and produces a fully resolved tree (e) reproduced from Pattengale et al. (2010) Uncovering hidden phylogenetic consensus.

Criticisms of Consensus Methods

- (1) consensus method lose character information—and therefore descriptive and explanatory power--relative to any one of the minimal length trees.
- (2) too much resolution--majority rule consensus trees indicate(monophyletic) groups not present in the set of best trees. (Something to keep in mind.)

Combined Data Approach

- = total evidence approach = analyze all data together if you can (and if it makes sense to do so)
- * many examples of morphology & molecular data sets where morphology has served an important role in determining topological relationships

Combined Data Approach

- * data set combination may yield more resolved tree than either data set alone, e.g., where one data set provides data for terminals and second data set provides characters for basal nodes.
- * may have weak but true signals in (both) data sets
- * even possible for combined data to conflict with consensus tree, e.g., in Barrett et al. (1991) where combined data tree is not found among any of the MPTs or consensus trees

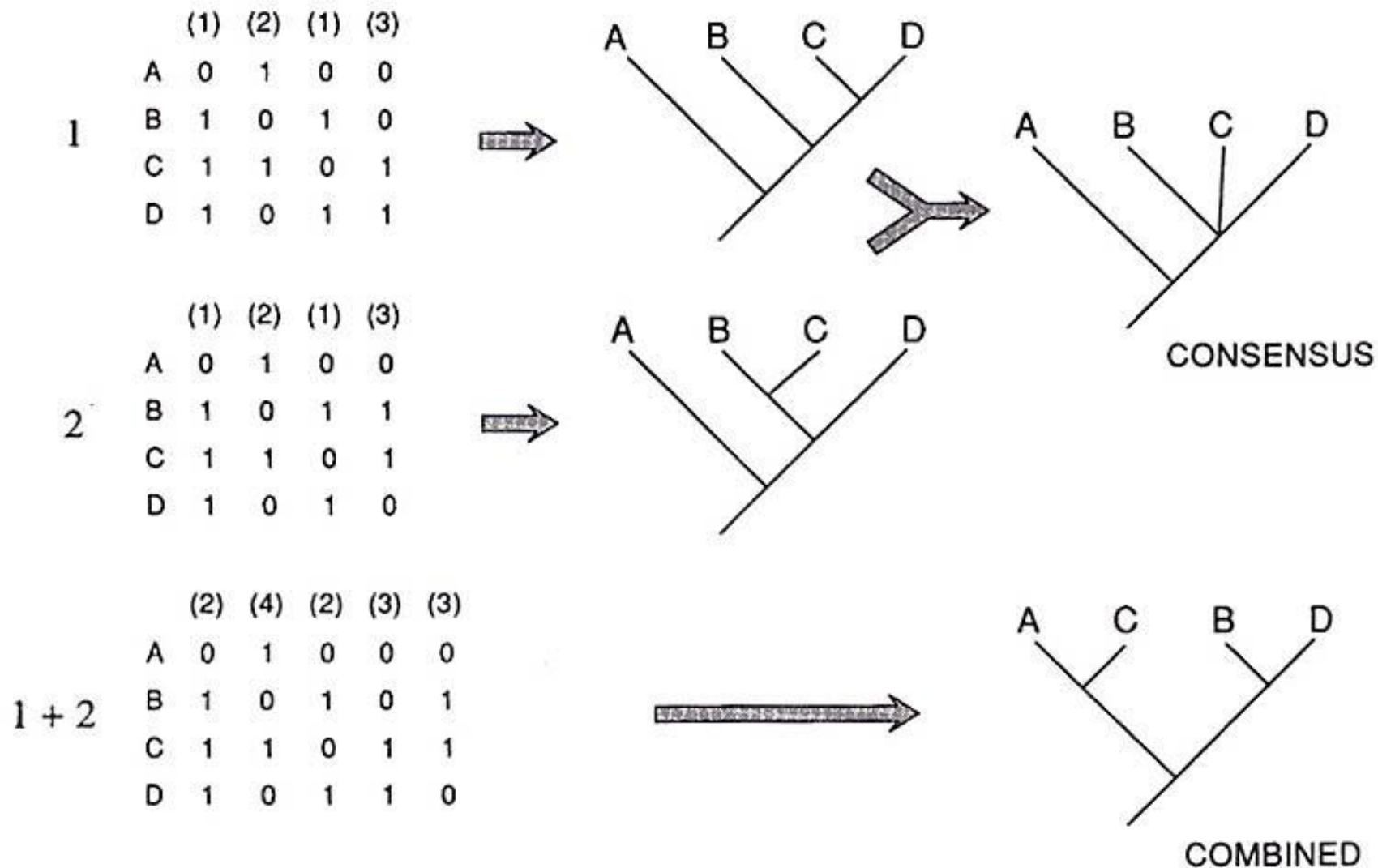


FIGURE 1. The consensus of the trees obtained from data sets 1 and 2 is incompatible with the tree obtained from the combined data set. See text for further explanation.

Types of Error in Phylogenetic Inference

Random Error: estimate of mean is not an accurate estimate of true population mean

- * most sites saturated
- * too small a sample size → estimate of true phylogeny is off
- * random error disappears with larger data sets

Systemic Error: where more data leads to more support for the wrong tree, i.e., it leads to inconsistency

- * non-independent substitutions (where characters/bases are under selection)
- * long branches (inconsistent for parsimony; problematic for all methods)
- * paralogy (different gene copies are being compared)
- * ancestral polymorphism (lineage sorting)
- * hybridization
- * horizontal gene transfer (xenology)

Concerted Evolution (A Paralogy Problem)

http://en.wikipedia.org/wiki/Concerted_evolution

Concerted evolution is a process that may explain the observation that [paralogous](#) genes within one species are more closely related to each other than to members of the same gene family in another species, even though the [gene duplication](#) event preceded the speciation event. The high sequence similarity between paralogs is maintained by homologous recombination events that lead to [gene conversion](#), effectively copying some sequence from one and overwriting the homologous region in the other.

An example can be seen in bacteria: *Escherichia coli* has seven operons encoding various Ribosomal RNA. For each of these genes, rDNA sequences are essentially identical among all of the seven operons (sequence divergence of only 0.195%). In a closely related species, *Haemophilus influenzae* its six ribosomal RNA operons are entirely identical. When the 2 species are compared together however, the sequence divergence of the 16S rRNA gene between them is 5.90%.^[1]



CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2007 February; 17(2): 184–191.
doi: [10.1101/gr.5457707](https://doi.org/10.1101/gr.5457707)

PMCID: PMC1781350

Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data

[Austen R.D. Ganley](#)^{1,3,4,5} and [Takehiko Kobayashi](#)^{1,2,4}

[Author information](#) ► [Article notes](#) ► [Copyright and License information](#) ►

This article has been [cited by](#) other articles in PMC.

Abstract

Go to:

Repeat families within genomes are often maintained with similar sequences. Traditionally, this has been explained by concerted evolution, where repeats in an array evolve “in concert” with the same sequence via continual turnover of repeats by recombination. Another form of evolution, birth-and-death evolution, can also explain this pattern, although in this case selection is the critical force maintaining the repeats. The level of intragenomic variation is the key difference between these two forms of evolution. The prohibitive size and repetitive nature of large repeat arrays have made determination of the absolute level of intragenomic repeat variability difficult, thus there is little evidence to support concerted evolution over birth-and-death evolution for many large repeat arrays. Here we use whole-genome shotgun sequence data from the genome projects of five fungal species to reveal absolute levels of sequence variation within the ribosomal RNA gene repeats (rDNA). The level of sequence variation is remarkably low. Furthermore, the polymorphisms that are detected are not functionally constrained and seem to exist beneath the level of selection. These results suggest the rDNA is evolving via concerted evolution. Comparisons with a repeat array undergoing birth-and-death evolution provide a clear contrast in the level of repeat array variation between these two forms of evolution, confirming that the rDNA indeed does evolve via concerted evolution. These low levels of intra-genomic variation are consistent with a model of concerted evolution in which homogenization is very rapid and efficiently maintains highly similar repeat arrays.

PubReader format:
click here to try

Formats:

[Article](#) | [PubReader](#) | [ePub \(beta\)](#) | [PDF \(404K\)](#)

Related citations in PubMed

- Monitoring the rate and dynamics of concerted evolution in the ribosomal DNA repeats of *Saccharomyces cerevisiae* [Mol Biol Evol. 2011]
- Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as [Genome Res. 2009]
- Diversity and recombination of dispersed ribosomal DNA and protein coding genes in microsporidia. [PLoS One. 2013]
- Short sequence repeats in microbial pathogenesis and evolution. [Cell Mol Life Sci. 1999]
- Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. [Trends Genet. 2006]

[See reviews...](#)

[See all...](#)

Cited by other articles in PMC

- Intra-Genomic Variation in the Ribosomal Repeats of Nematodes [PLoS ONE.]
- Metschnikowia Species Share a Pool of Diverse rRNA Genes Differing in Regions That Determine Hairpin-Loop S [PLoS ONE.]
- Intragenomic Profiling Using Multicopy Genes: The rDNA Internal Transcribed Spacer Sequences of the Freshwater *Hydra* [PLoS ONE.]
- Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribe [Ecology and Evolution. 2013]
- Molecular diagnosis to discriminate pathogen and apathogen species of the hybrid *Verticillium* [Applied Microbiology and Biotech...]

[See all...](#)

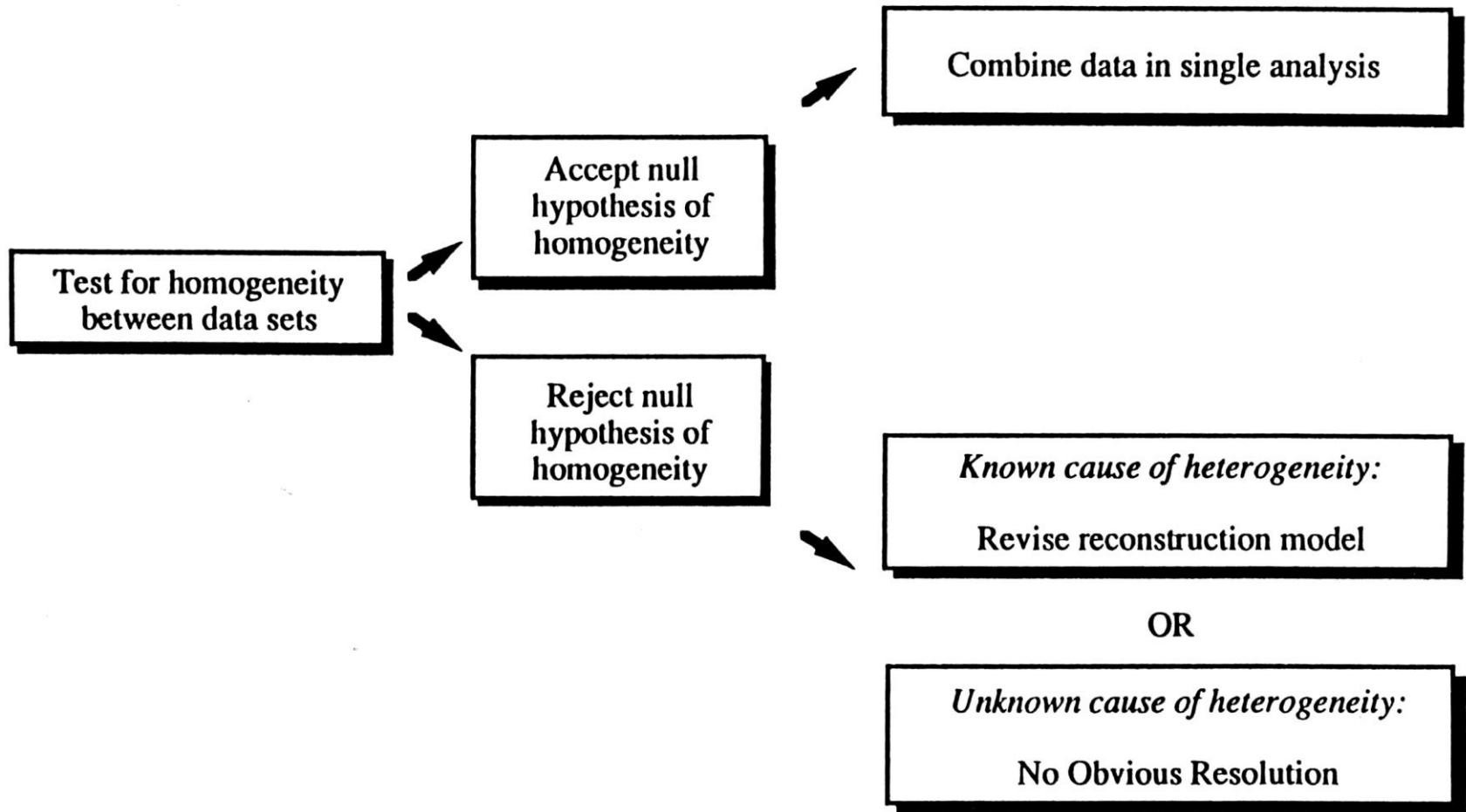
Links

When to Combine and When to Use Consensus

- * If there is error in your data set then watch out for combined data approaches
- * If one data set is providing wrong signal, isn't it better to have one right and one wrong answer from separate analyses, than a single incorrect one?
- * If different trees result from different data sets and each is strongly supported (earmarks of systematic error) use consensus methods that only accept groups found in set of best trees.

Data Partitions and Congruence

- * Issues of data combination and consensus are also relevant to single data sets
- * Think about different data partitions of a single data set:
 - e.g. different genes
 - e.g. process partitions
 - coding and non-coding sections of a gene
 - 1st, 2nd, and 3rd positions
 - non-synonymous vs synonymous nucleotide changes
 - different classes of amino acid substitutions
- * Bull et al. (1993) recommend doing *homogeneity tests* to look for data set agreement



Comparing Trees

There are many circumstances when we need to compare trees:

- * choosing best model of nucleotide substitution (*nested* models)
- * testing tree against favored/alternative hypothesis (topology) (e.g., the existing classification)
- * testing goodness of best vs. suboptimal trees
- * in likelihood and Bayesian frameworks comparing branch length estimates
- * comparing combinability of trees (homogeneity tests)
- * comparing trees without the intention of combining:
e.g., comparing tree structures to biogeographic pattern,
or for evidence of co-cladogenesis (co-speciation)
- * etc.

Incongruence Length Difference Test

Farris et al. (1994)

(tests whether data partitions can be combined in parsimony analysis)

$$D_{XY} = L_{(X+Y)} - (L_X + L_Y)$$

where $L_{(X+Y)}$ = length of tree from combined data

L_X = length of tree from data set X

L_Y = length of tree from data set Y

D_{XY} = measure of incongruence

D_{XY} = is large, when combining data creates substantial (additional) homoplasy

$L_{(X+Y)}$ exceeds $(L_X + L_Y)$ by the amount of extra homoplasy required by the combined approach.

Incongruence Length Difference Test

- * A significance test of the data partition incongruence can be generated by comparing the combined data tree length to values generated by random partitionings of the data sets.
- * 100 to 1000 random partitions (P_P , P_Q) are made by generating data sets (matrices) identical in size to the original data sets (but now with random mixes of characters from the two original data sets)
- * tree lengths (L_P , L_Q), for random partitions are calculated for each partition
- * because the structure (signal) in the original partitions is randomized the new partitions and new (randomized) trees may get longer (esp. if the partitions are incongruent)
- * but combining them will no longer add much additional homoplasy
- * so D_{PQ} values will be small
- * compare the number of times $D_{XY} > D_{PQ}$
- * if D_{XY} is greater than D_{PQ} more than 95% of the time do not combine partitions

Incongruence Length Difference Test

- * implemented in PAUP as Partition Homogeneity Test
- * sensitive to invariant sites and inclusion of autapomorphies
- * not a indicator of data set congruence, if evolutionary rates for the data sets are different (Barker and Lutzoni 2002 and Darlu and Lecointre 2002)
- * even with significant partition heterogeneity, often possible to combine data and get improved phylogenetic accuracy (Barker and Lutzoni, 2002)

Data Partitions

- there is no set rule for number
- sometimes partition assigned for each gene
- even more important is to assign a partition for each codon position
- or both of the above
- the biggest advantage to partitioning is that each partition can be modeled separately i.e., one can employ a different evolutionary model: e.g., morphology versus individual genes (e.g., in MrBayes)

Model Tests (optional)

- hierarchical likelihood ratio test
- AIC – Akaike Information Criterion (measures loss of information when wrong model is used)
- Bayesian Test of Models – can employ Bayes factors, posterior probabilities, or Bayes Information Criterion
- Decision theory methods
- jModelTest and jModelTest2 (Posada et al. 2012) see (<https://code.google.com/p/jmodeltest2/>)
- MrModeltest (for MrBayes Vers. 3) (Nylander 2008) (<http://www.abc.se/~nylander/mrmodeltest2/mrmodeltest2.html>)

See Posada, D. 2009. Selecting models of evolution. pp. 345-361. *In* The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis. P. Lemey, M. Salemi, and A Vandamme (eds.). Cambridge Univ. Press.

Hierarchical Likelihood Ratio Test

- * use when models are nested (i.e., one case is a special case of the other)

- * hierarchical likelihood ratio test

$$\text{hLRT} = 2(\ln L_1 - \ln L_2)$$

where $\ln L_1$ maximum log likelihood under more parameter rich model

and $\ln L_2$ maximum log likelihood under less parameter-rich model

- * essentially a X^2 distribution (with degrees of freedom equal to number of extra parameters in more complex model)
- * use $-\ln L_1$ (more complex model) when LRT is sufficiently large

Consider two hierarchically nested substitution models: HKY85 and GTR. The GTR model differs from HKY85 by the addition of four additional rate parameters. Imagine we had the likelihood scores of the two models for a neighbor-joining tree:

HKY85 $-\ln L = 1787.08$

GTR $-\ln L = 1784.82$

Then, $LRT = 2 (1787.08 - 1784.82) = 4.53$

$df = 4$ (GTR adds 4 additional parameters to HKY85)

critical value ($P = 0.05$) = 9.49

In this case, GTR does not fit the data significantly better than HKY85, and we infer that the four rate additional rate parameters are not biologically meaningful (given our power to detect such differences).

[source: http://www.molecularrevolution.org/resources/lrt](http://www.molecularrevolution.org/resources/lrt)

Maximum likelihood inference(AIC) (Akaike, 1974)

- * $AIC = -2\ln L + 2n$
- * Where, $\ln L$ is the maximum likelihood value of a specific model of nucleotide sequence evolution and tree topology given the data.
- * n = the number of parameters free to vary
- * Smaller AIC indicates a better model

source: <http://bio.fsu.edu/~stevet/BSC5936/Wilgenbusch.2003.pdf>

Comparing Tree Topologies (lab)

- * Kishino-Hasegawa (KH) test
- * Shimodaira-Hasegawa test
- * Weighted test variants
- * Approximately unbiased (AU) test
- * Swofford-Olsen-Waddell-Hillis (SOWH)

- * Software:
Consel: <http://www.is.titech.ac.jp/~shimo/prog/consel/>
Tree Puzzle: <http://www.tree-puzzle.de/>

See Schmidt 2009. Testing Tree topologies. pp. 381-496. *In* The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis. P. Lemey, M. Salemi, and A Vandamme (eds.). Cambridge Univ. Press.

Tree comparing software: Consel and Tree Puzzle

CONSEL home page


www.is.titech.ac.jp/~shimo/prog/consel/

CONSEL HOME PAGE

CONSEL: for assessing the confidence of phylogenetic tree selection

25,756 Visitors

27 Feb 2010 - 15 Oct 2011



WHAT IS CONSEL?

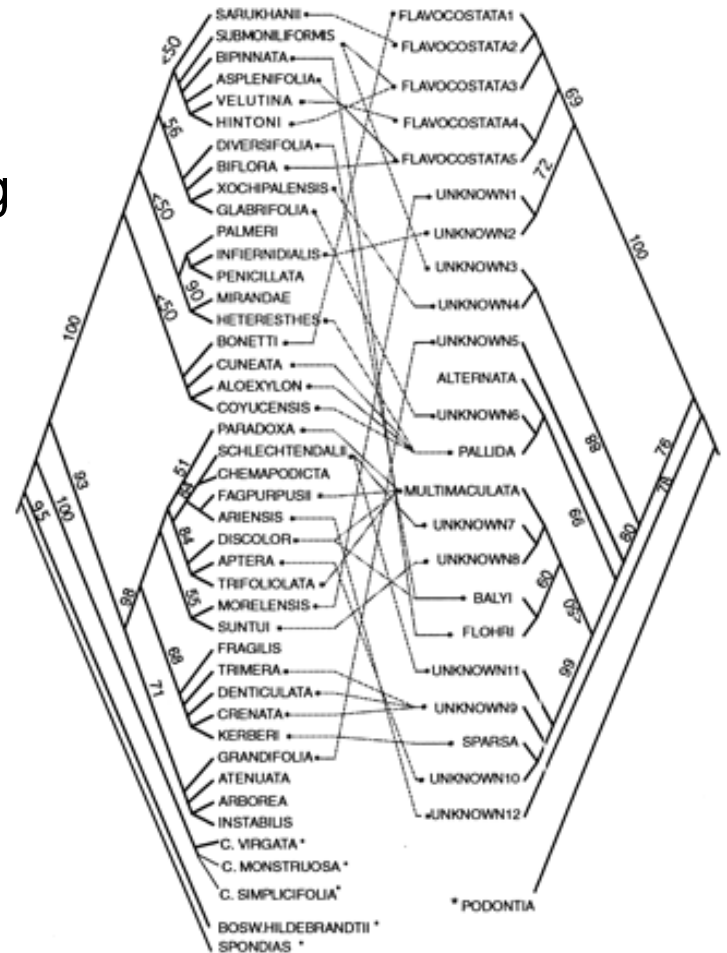
CONSEL is a program package consists of small programs written in C language. It calculates the probability value (i.e., p-value) to assess the confidence in the selection problem. Although CONSEL is applicable to any selection problem, it is mainly designed for the phylogenetic tree selection. CONSEL does not estimate the phylogenetic tree by itself, but CONSEL does read the output of the other phylogenetic packages, such as *Molphy*, *PAML*, *PAUP**, *TREE-PUZZLE*, and *PhyML*. CONSEL calculates the p-value using several testing procedures; the bootstrap probability, the Kishino-Hasegawa test, the Shimodaira-Hasegawa test, and the weighted Shimodaira-Hasegawa test. In addition to these conventional tests, CONSEL calculates the p-value based on the approximately unbiased test using the multi-scale bootstrap technique. This newly developed method gives less biased results than the conventional methods.

QUICK GUIDE: [Using CONSEL is Easy!](#) (updated 2008/09/12 for PAUP; 2010/01/30 for PhyML)

Quantifying Incongruence

Sometimes we have no intention of combining the trees and simply want to know to what extent our trees agree or disagree. For example:

1. You have a tree and a biogeographic pattern and want to know how well they fit or how much they disagree?
2. You have host and parasite and want to quantify degree of fit/mismatch to evaluate evidence for co-speciation?



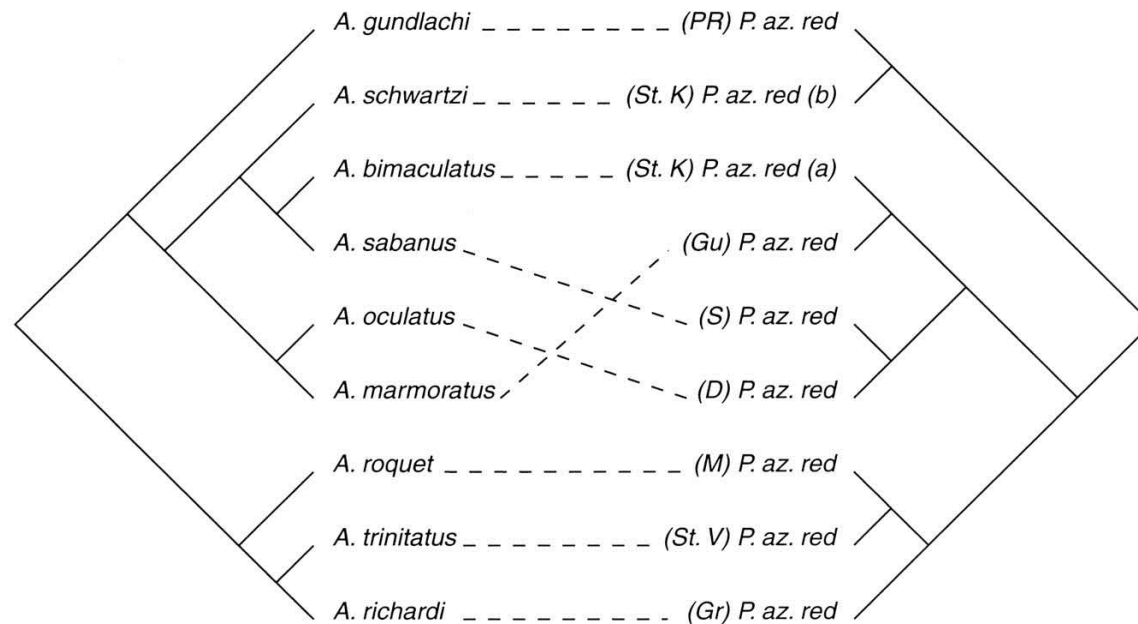
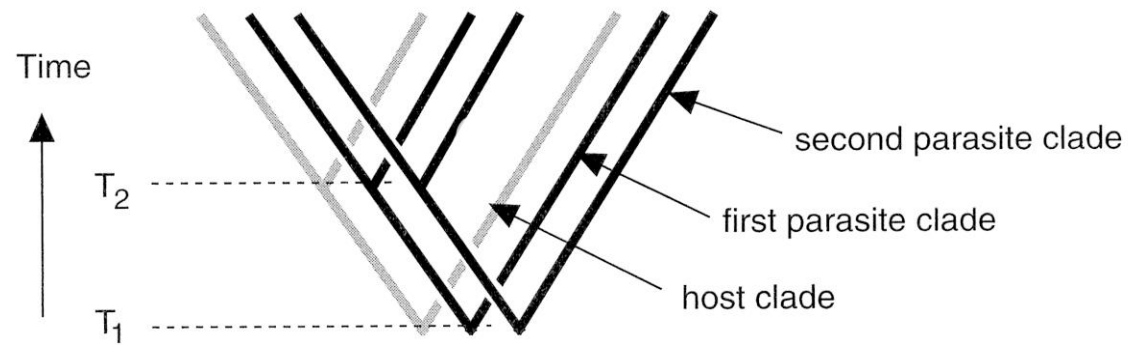
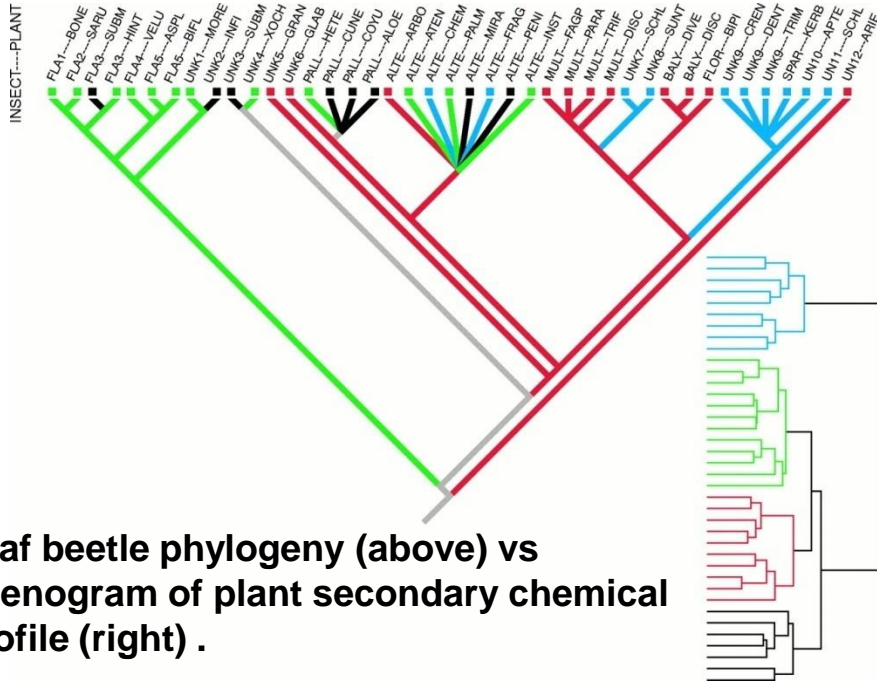
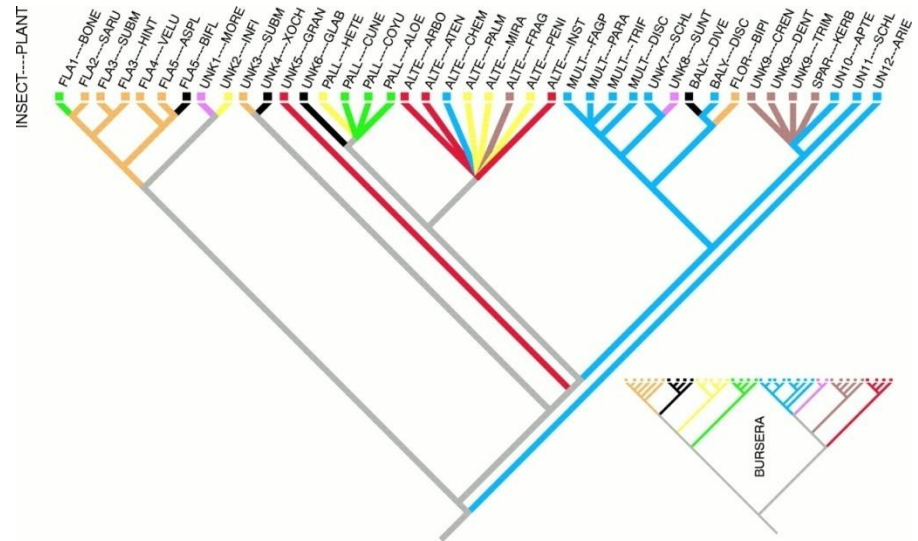


FIGURE 3.12. Relationships between Caribbean *Anolis* and *Plasmodium azurophilum* red (host tree adapted from Roughgarden, 1995, with parasite tree and host/parasite associations from Perkins, 2000).

Insects on Plants: Macroevolutionary Chemical Trends in Host Use



Becerra. 1997. Science 276: 253-256



Leaf beetle phylogeny (above) vs hostplant phylogeny (right)