

Introduction to characters and parsimony analysis

Genetic Relationships

- **Genetic relationships exist between individuals within populations**
- **These include ancestor-descendent relationships and more indirect relationships based on common ancestry**
- **Within sexually reducing populations there is a network of relationships**
- **Genetic relations within populations can be measured with a coefficient of genetic relatedness**

Phylogenetic Relationships

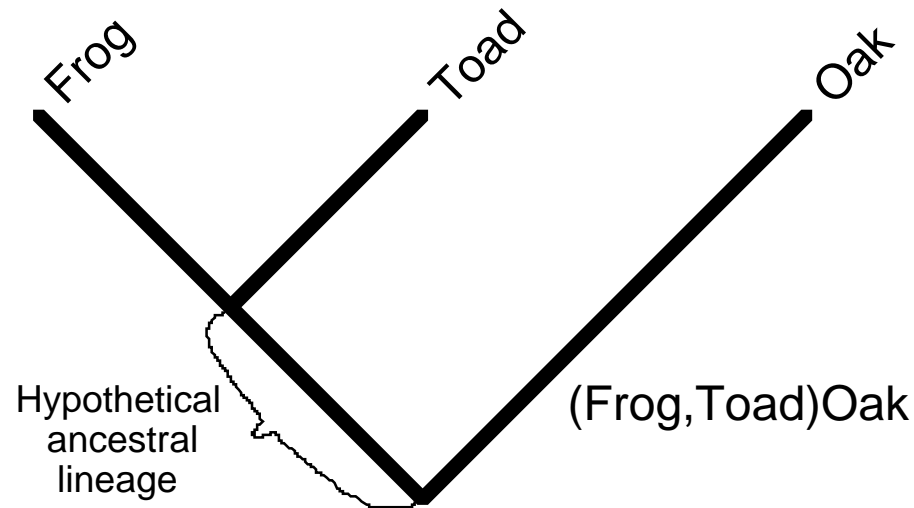
- **Phylogenetic relationships exist between lineages (e.g. species, genes)**
- **These include ancestor-descendent relationships and more indirect relationships based on common ancestry**
- **Phylogenetic relationships between species or lineages are (expected to be) tree-like**
- **Phylogenetic relationships are not measured with a simple coefficient**

Phylogenetic Relationships

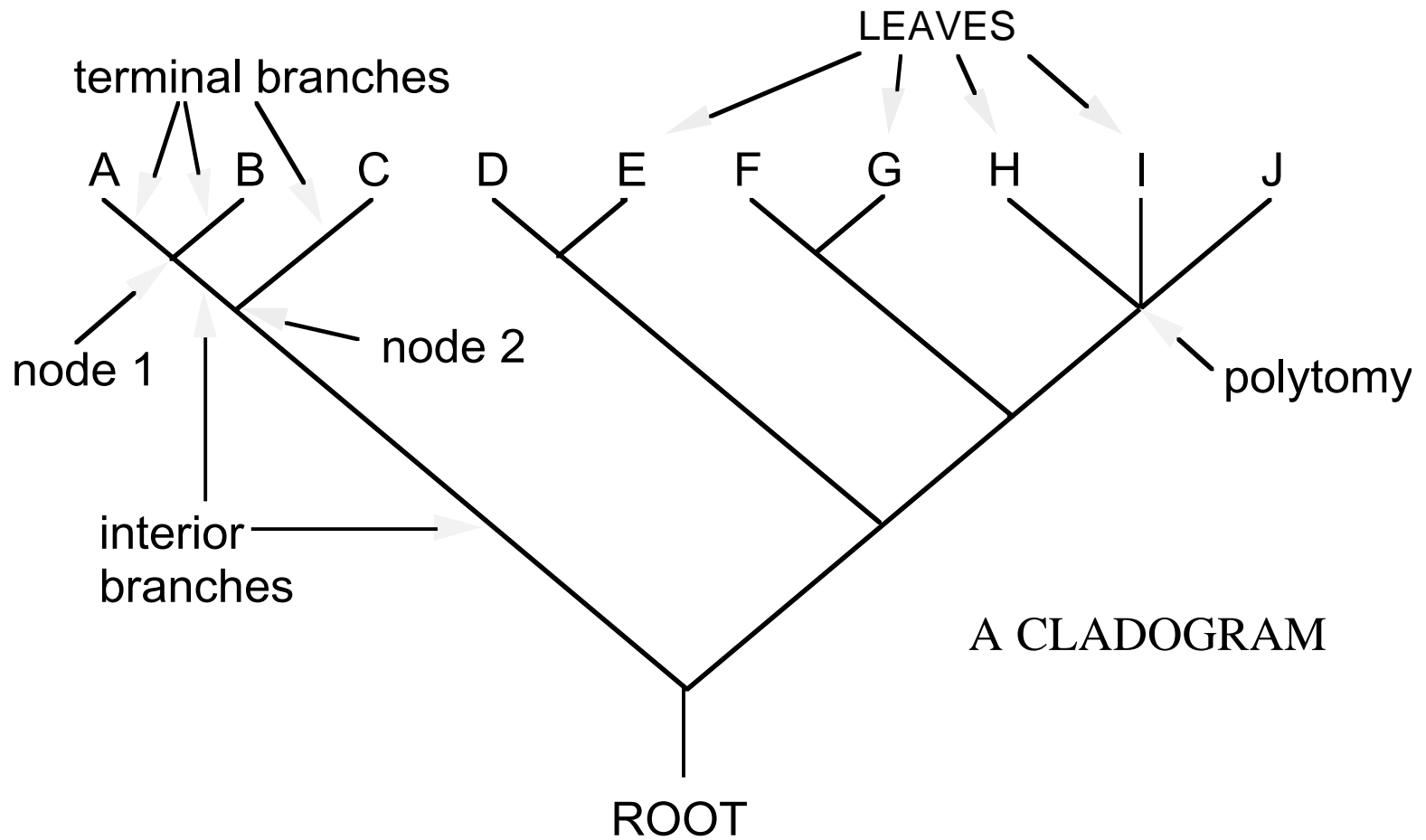
- **Traditionally phylogeny reconstruction was dominated by the search for ancestors, and ancestor-descendant relationships**
- **In modern phylogenetics there is an emphasis on indirect relationships**
- **Given that all lineages are related, closeness of phylogenetic relationships is a relative concept.**

Phylogenetic relationships

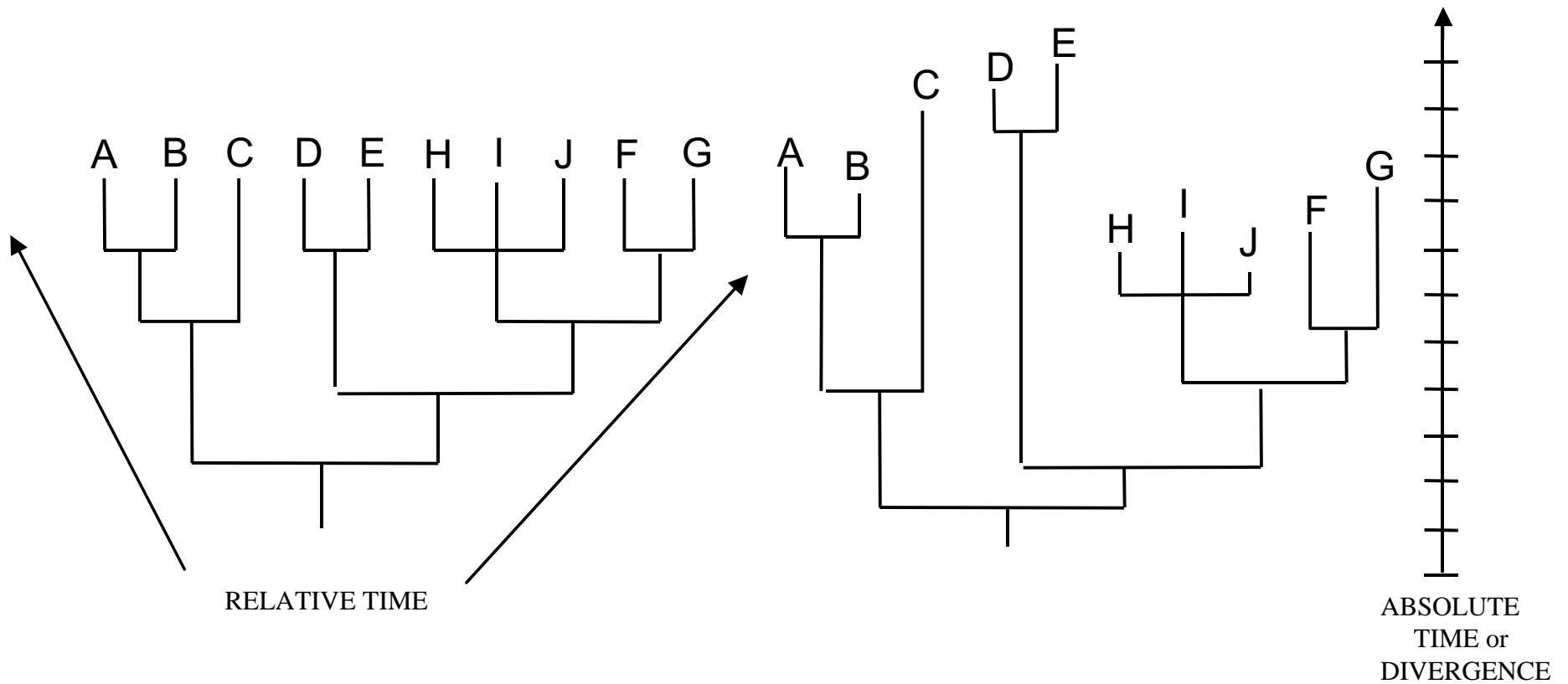
- **Two lineages are more closely related to each other than to some other lineage if they share a more recent common ancestor - this is the cladistic concept of relationships**
- **Phylogenetic hypotheses are hypotheses of common ancestry**



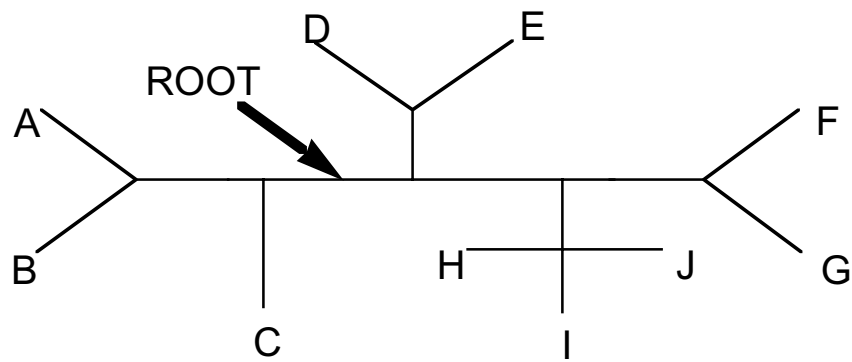
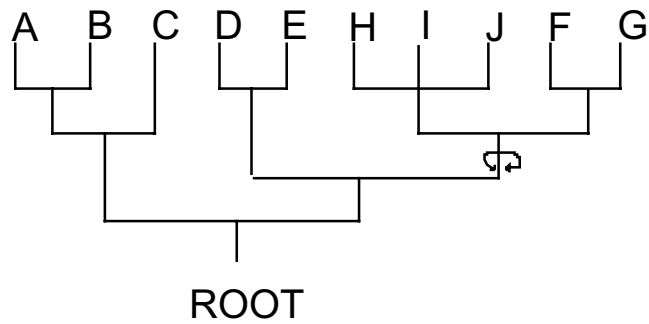
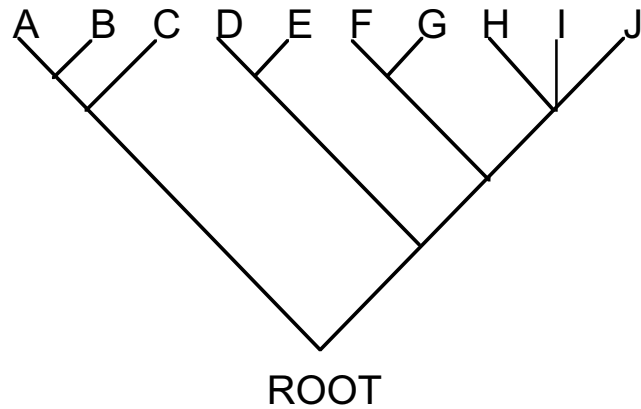
Phylogenetic Trees



CLADOGRAMS AND PHYLOGRAMS



Trees - Rooted and Unrooted



Characters and Character States

- **Organisms comprise sets of features**
- **When organisms/taxa differ with respect to a feature (e.g. its presence or absence or different nucleotide bases at specific sites in a sequence) the different conditions are called *character states***
- **The collection of character states with respect to a feature constitute a *character***

Character evolution

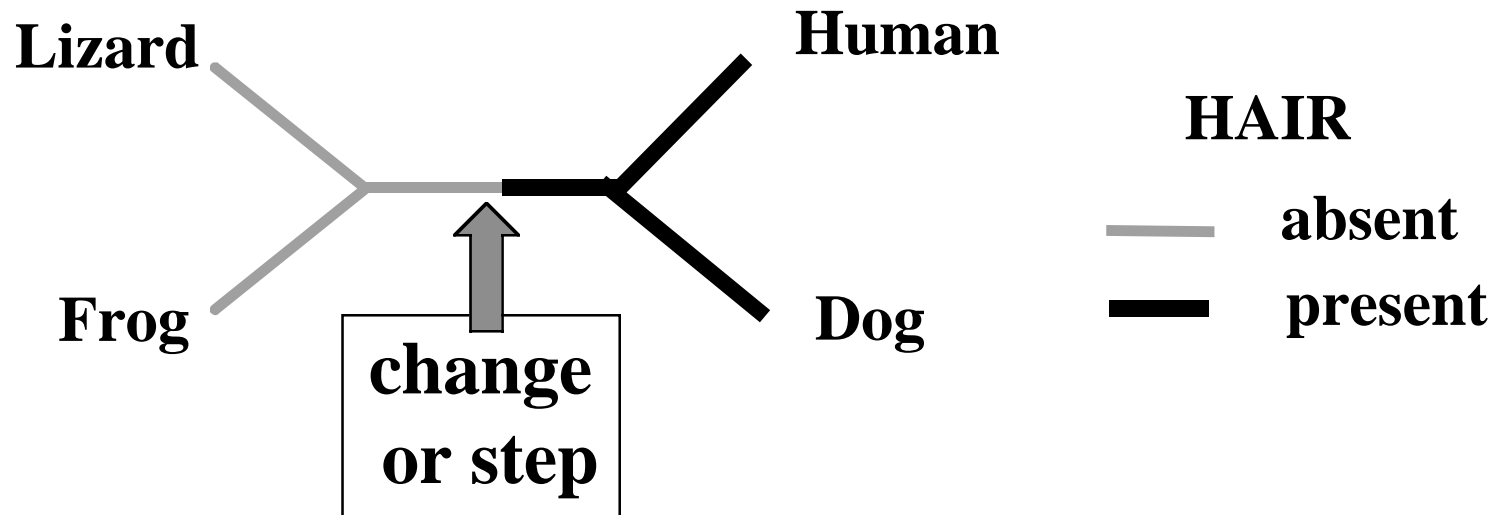
- **Heritable changes (in morphology, gene sequences, etc.) produce different character states**
- **Similarities and differences in character states provide the basis for inferring phylogeny (i.e. provide evidence of relationships)**
- **The utility of this evidence depends on how often the evolutionary changes that produce the different character states occur independently**

Unique and unreversed characters

- **Given a heritable evolutionary change that is unique and unreversed (e.g. the origin of hair) in an ancestral species, the presence of the novel character state in any taxa must be due to inheritance from the ancestor**
- **Similarly, absence in any taxa must be because the taxa are not descendants of that ancestor**
- **The novelty is a *homology* acting as badge or marker for the descendants of the ancestor**
- **The taxa with the novelty are a clade (e.g. Mammalia)**

Unique and unreversed characters

- Because hair evolved only once and is unreversed (not subsequently lost) it is *homologous* and provides unambiguous evidence for relationships

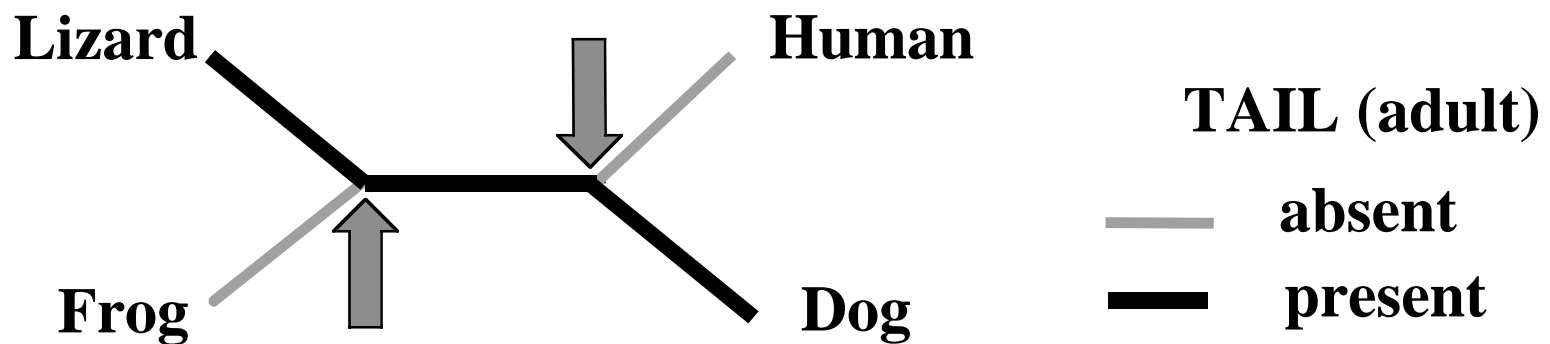


Homoplasy - Independent evolution

- **Homoplasy is similarity that is not homologous (not due to common ancestry)**
- **It is the result of independent evolution (convergence, parallelism, reversal)**
- **Homoplasy can provide misleading evidence of phylogenetic relationships (if mistakenly interpreted as homology)**

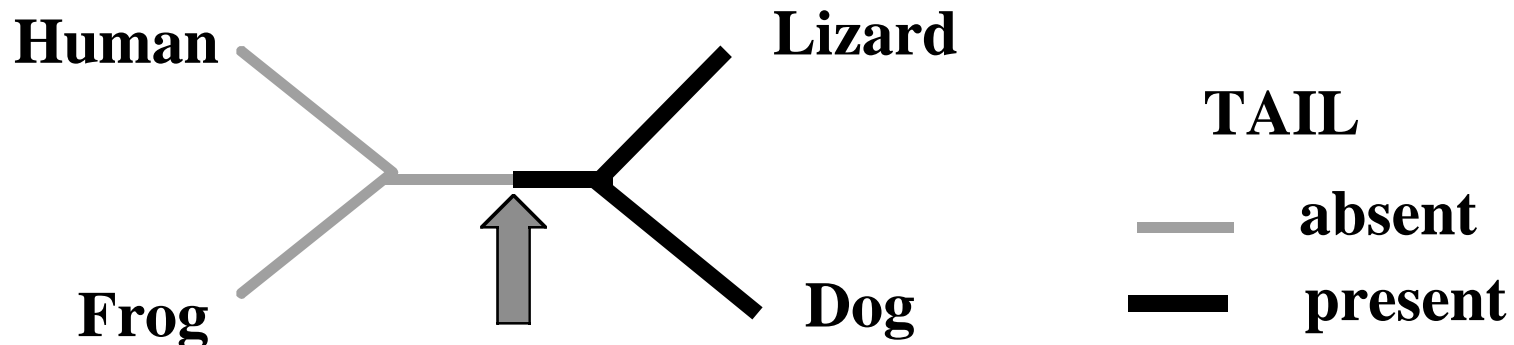
Homoplasy - independent evolution

- **Loss of tails evolved independently in humans and frogs - there are two steps on the true tree**



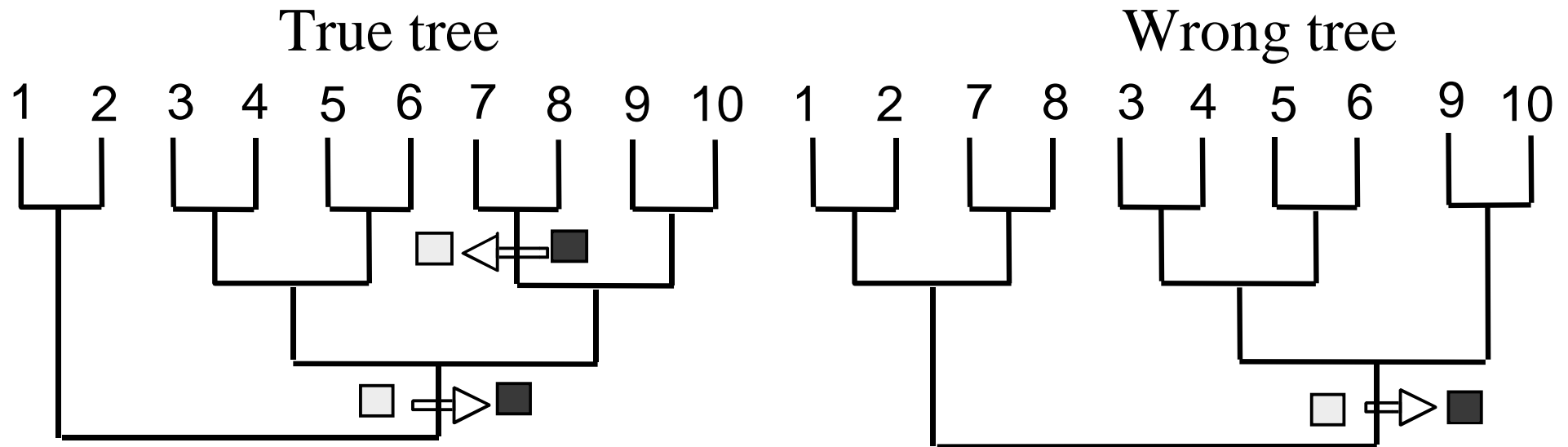
Homoplasy - misleading evidence of phylogeny

- If misinterpreted as homology, the absence of tails would be evidence for a wrong tree: grouping humans with frogs and lizards with dogs



Homoplasy - reversal

- **Reversals are evolutionary changes back to an ancestral condition**
- **As with any homoplasy, reversals can provide misleading evidence of relationships**



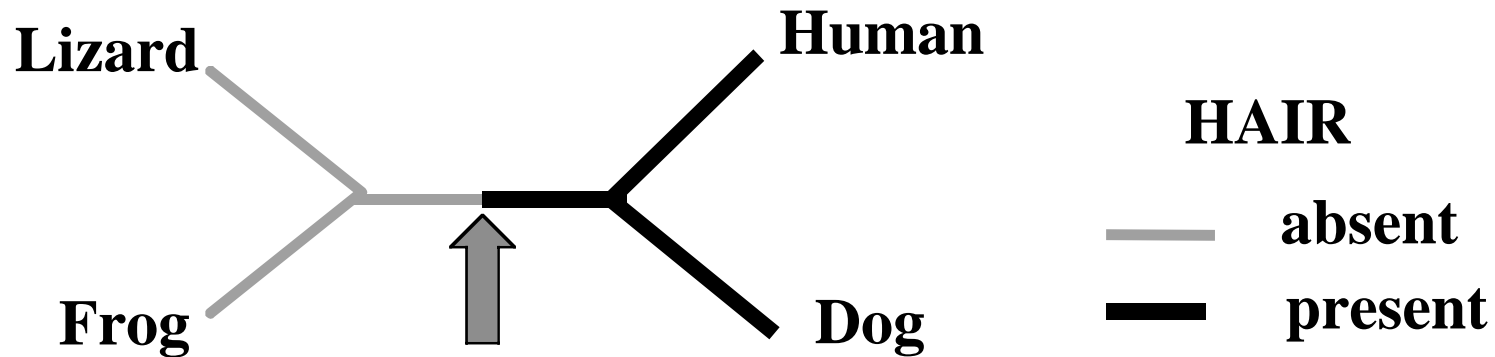
Homoplasy - a fundamental problem of phylogenetic inference

- If there were no homoplastic similarities inferring phylogeny would be easy - all the pieces of the jig-saw would fit together neatly**
- Distinguishing the misleading evidence of homoplasy from the reliable evidence of homology is a fundamental problem of phylogenetic inference**

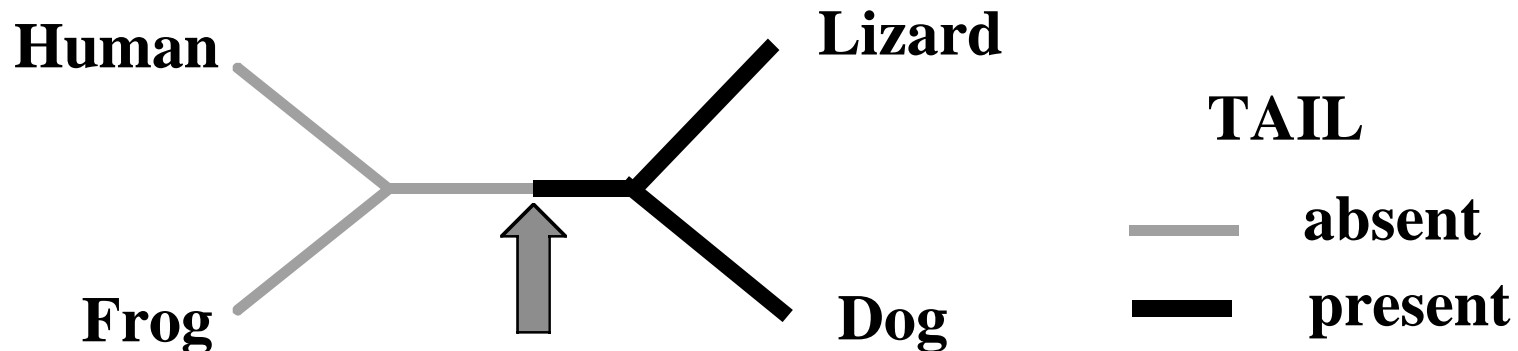
Homoplasy and Incongruence

- If we assume that there is a single correct phylogenetic tree then:**
- When characters support conflicting phylogenetic trees we know that there must be some misleading evidence of relationships among the incongruent or incompatible characters**
- Incongruence between two characters implies that at least one of the characters is homoplastic and that at least one of the trees the character supports is wrong**

Incongruence or Incompatibility



- These trees and characters are incongruent - both trees cannot be correct, at least one is wrong and at least one character must be homoplastic



Distinguishing homology and homoplasy

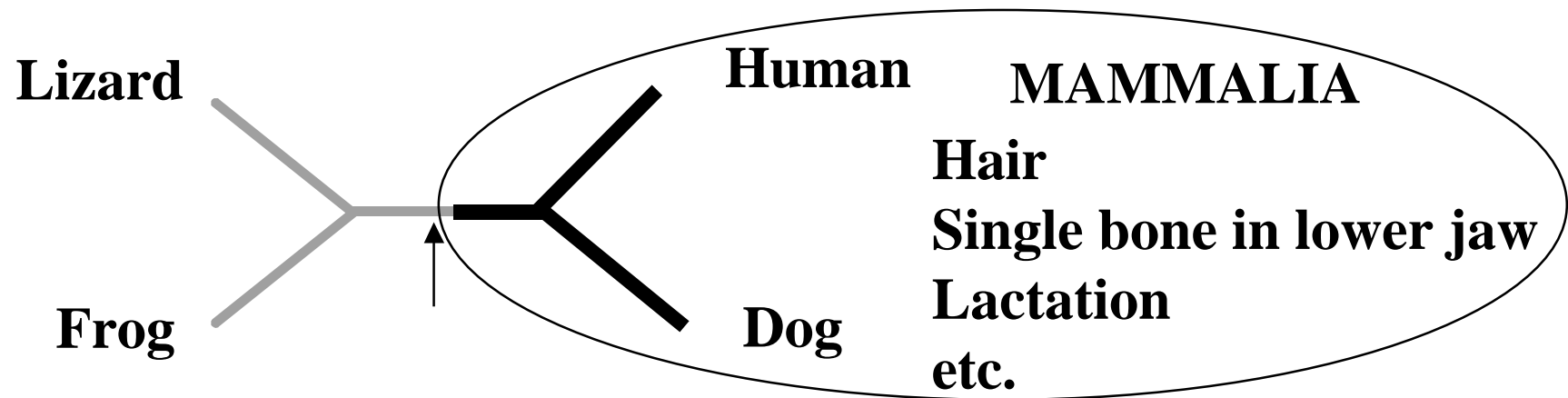
- **Morphologists use a variety of techniques to distinguish homoplasy and homology**
- **Homologous features are expected to display detailed similarity (in position, structure, development) whereas homoplastic similarities are more likely to be superficial**
- **As recognised by Charles Darwin congruence with other characters provides the most compelling evidence for homology**

The importance of congruence

- **“The importance, for classification, of trifling characters, mainly depends on their being correlated with several other characters of more or less importance. The value indeed of an aggregate of characters is very evident a classification founded on any single character, however important that may be, has always failed.”**
- **Charles Darwin: Origin of Species, Ch. 13**

Congruence

- We prefer the 'true' tree because it is supported by multiple congruent characters



Homoplasy in molecular data

Incongruence and therefore homoplasy can be common in molecular sequence data

- There are a limited number of alternative character states (e.g. Only A, G, C and T in DNA)**
- Rates of evolution are sometimes high**

Character states are chemically identical

- homology and homoplasy are equally similar**
- cannot be distinguished by detailed study of similarity and differences**

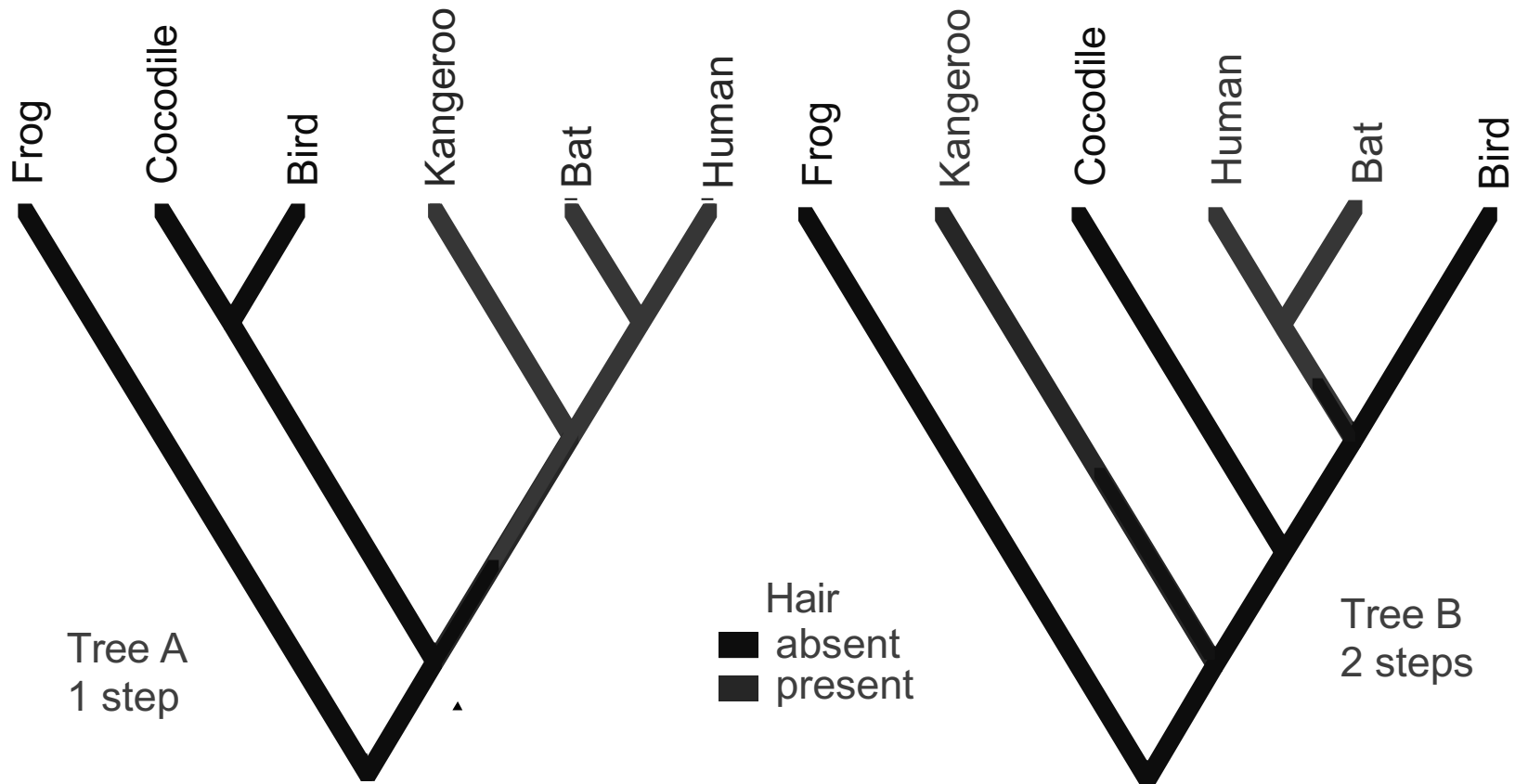
Parsimony analysis

- **Parsimony methods provide one way of choosing among alternative phylogenetic hypotheses**
- **The parsimony criterion favours hypotheses that maximise congruence and minimise homoplasy**
- **It depends on the idea of the fit of a character to a tree**

Character Fit

- **Initially, we can define the fit of a character to a tree as the minimum number of steps required to explain the observed distribution of character states among taxa**
- **This is determined by parsimonious character optimization**
- **Characters differ in their fit to different trees**

Character Fit

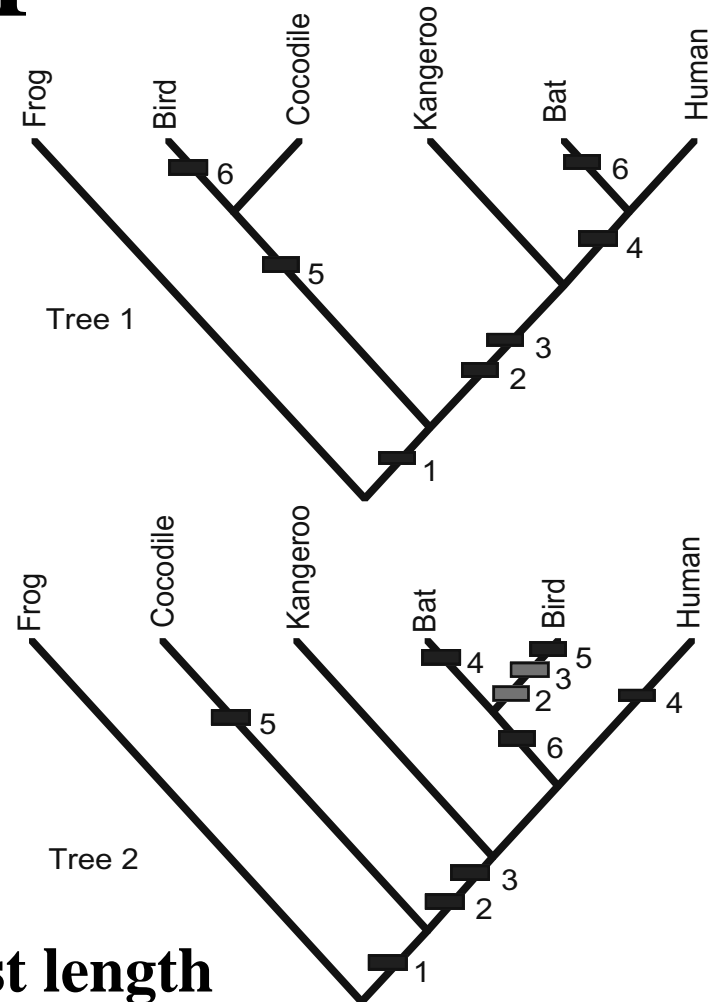


Parsimony Analysis

- **Given a set of characters, such as aligned sequences, parsimony analysis works by determining the fit (number of steps) of each character on a given tree**
- **The sum over all characters is called Tree Length**
- **Most parsimonious trees (MPTs) have the minimum tree length needed to explain the observed distributions of all the characters**

Parsimony in practice

		CHARACTERS						
		1	2	3	4	5	6	
		amnion	hair	lactation	placenta	antorbital fenestra	wings	
TAXA	Frog	-	-	-	-	-	-	
	Bird	+	-	-	-	+	+	
	Crocodile	+	-	-	-	+	-	
	Kangaroo	+	+	+	-	-	-	
	Bat	+	+	+	+	-	+	
	Human	+	+	+	+	-	-	TREE LENGTH
FIT	Tree 1	1	1	1	1	1	2	7
	Tree 2	1	2	2	2	2	1	10



Of these two trees, Tree 1 has the shortest length and is the most parsimonious
Both trees require some homoplasy (extra steps)

Results of parsimony analysis

- **One or more most parsimonious trees**
- **Hypotheses of character evolution associated with each tree (where and how changes have occurred)**
- **Branch lengths (amounts of change associated with branches)**
- **Various tree and character statistics describing the fit between tree and data**
- **Suboptimal trees - optional**

Character types

- Characters may differ in the costs (contribution to tree length) made by different kinds of changes

- Wagner (ordered, additive)

0 — 1 — 2 (morphology, unequal costs)

- Fitch (unordered, non-additive)

A — G (morphology, molecules)

T — C

(equal costs for all changes)

— one step
— two steps

Character types

- **Sankoff (generalised)**

A—G (morphology, molecules)



T—C (user specified costs)

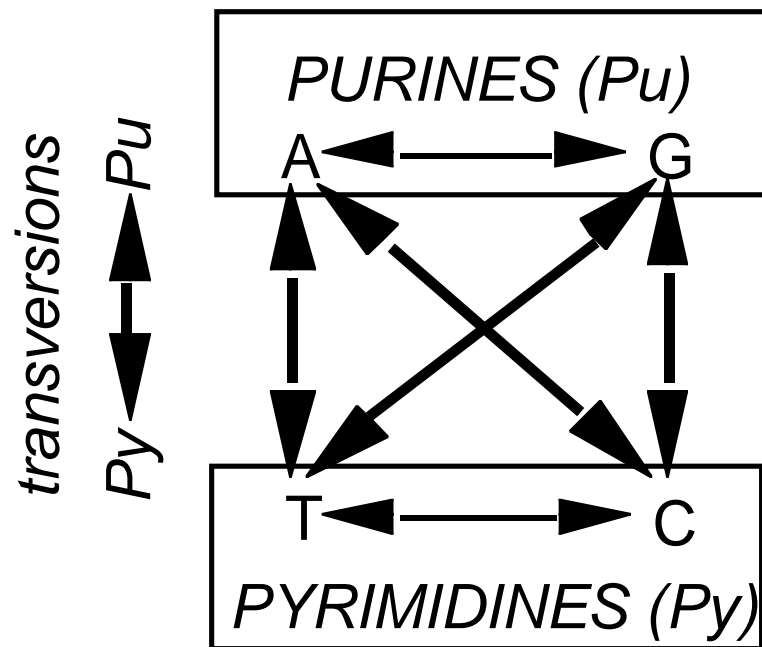
— one step

— five steps

- **For example, differential weighting of transitions and transversions**
- **Costs are specified in a stepmatrix**
- **Costs are usually symmetric but can be asymmetric also (e.g. costs more to gain than to loose a restriction site)**

Stepmatrices

- **Stepmatrices specify the costs of changes within a character**



		To			
		A	C	G	T
From	A	0	5	1	5
	C	5	0	5	1
	G	1	5	0	5
	T	5	1	5	0

transitions
Py —▶ *Py*
Pu —▶ *Pu*

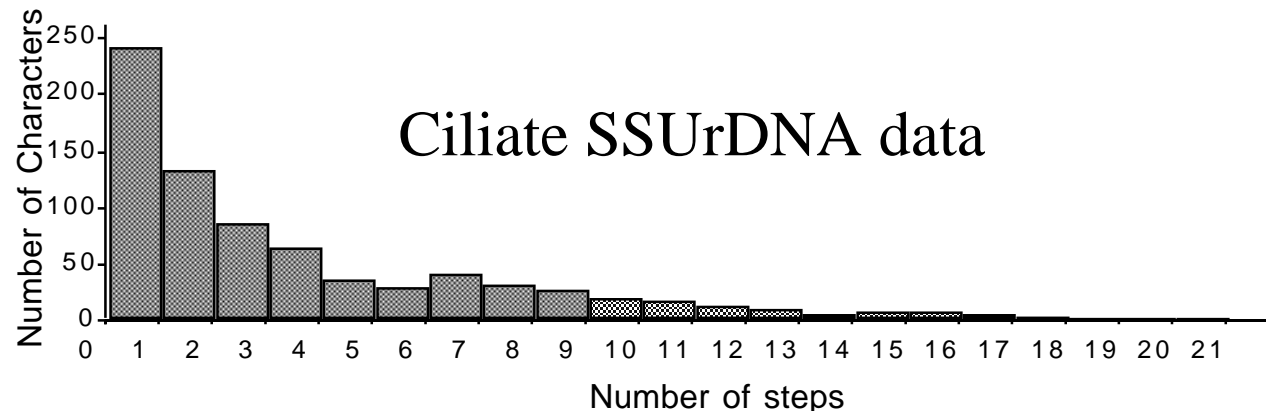
Different characters (e.g 1st, 2nd and 3rd codon positions can also have different weights

Weighted parsimony

- **If all kinds of steps of all characters have equal weight then parsimony:**
 - **Minimises homoplasy (extra steps)**
 - **Maximises the amount of similarity due to common ancestry**
 - **Minimises tree length**
- **If steps are weighted unequally parsimony minimises tree length - a weighted sum of the cost of each character**

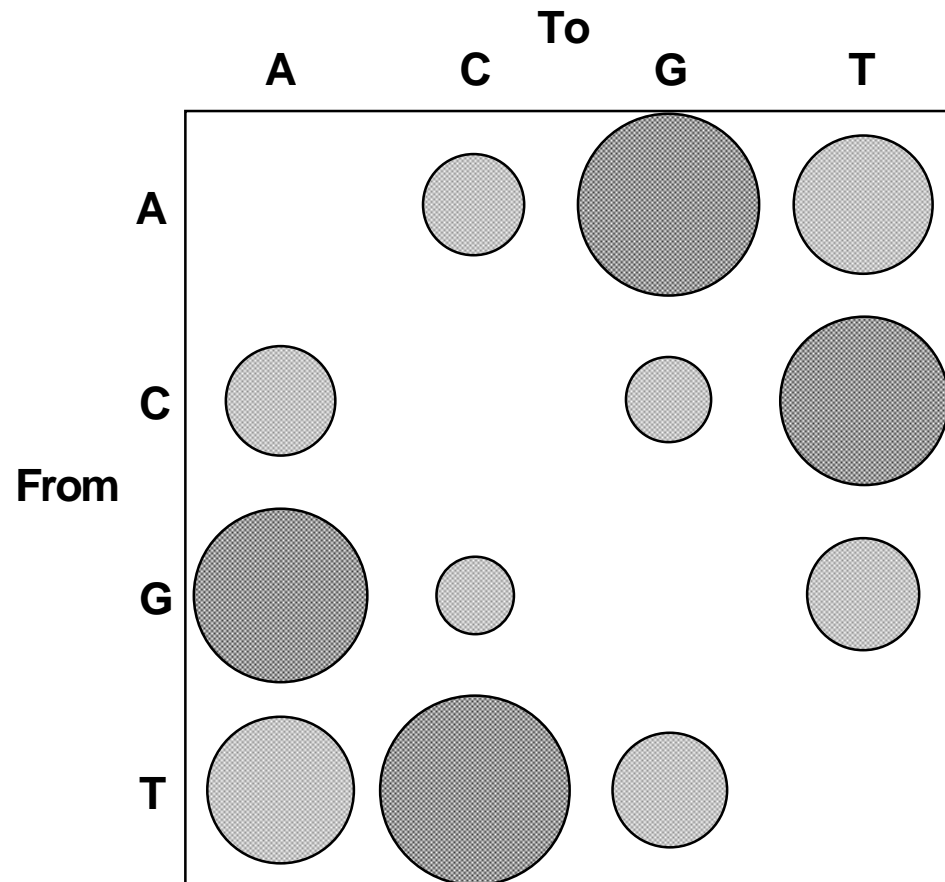
Why weight characters?

- Many systematists consider weighting unacceptable, but weighting is unavoidable (unweighted = equal weights)
- Transitions may be more common than transversions
- Different kinds of transitions and transversions may be more or less common
- Rates of change may vary with codon positions
- The fit of different characters on trees may indicate differences in their reliabilities



- However, equal weighting is the commonest procedure and is the simplest (but probably not the best) approach

Different kinds of changes differ in their frequencies



● Transitions
○ Transversions

**Unambiguous changes
on most parsimonious
tree of Ciliate SSUrDNA**

Parsimony - advantages

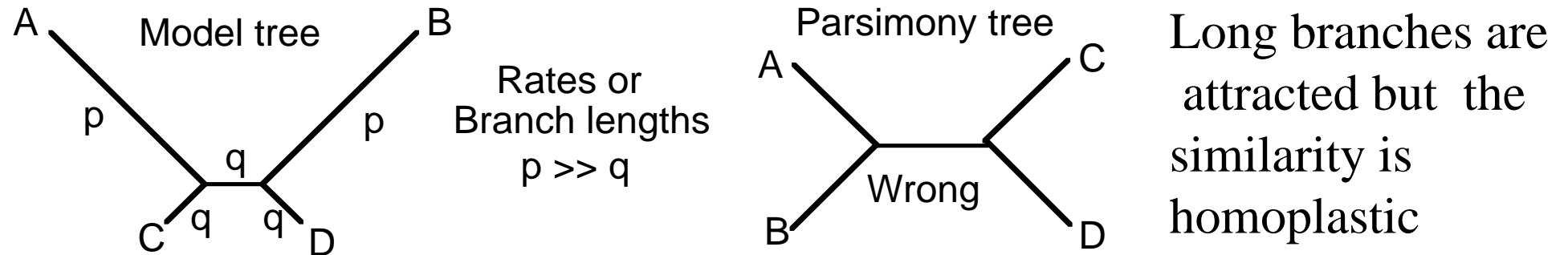
- **is a simple method - easily understood operation**
- **does not seem to depend on an explicit model of evolution**
- **gives both trees and associated hypotheses of character evolution**
- **should give reliable results if the data is well structured and homoplasy is either rare or widely (randomly) distributed on the tree**

Parsimony - disadvantages

- **May give misleading results if homoplasy is common or concentrated in particular parts of the tree, e.g:**
 - **thermophilic convergence**
 - **base composition biases**
 - **long branch attraction**
- **Underestimates branch lengths**
- **Model of evolution is implicit - behaviour of method not well understood**
- **Parsimony often justified on purely philosophical grounds - we must prefer simplest hypotheses - particularly by morphologists**
- **For most molecular systematists this is unconvincing**

Parsimony can be inconsistent

- Felsenstein (1978) developed a simple model phylogeny including four taxa and a mixture of short and long branches
- Under this model parsimony will give the wrong tree



- With more data the certainty that parsimony will give the wrong tree increases - so that parsimony is statistically inconsistent
- Advocates of parsimony initially responded by claiming that Felsenstein's result showed only that his model was unrealistic
- It is now recognised that the long-branch attraction (in the Felsenstein Zone) is one of the most serious problems in phylogenetic inference

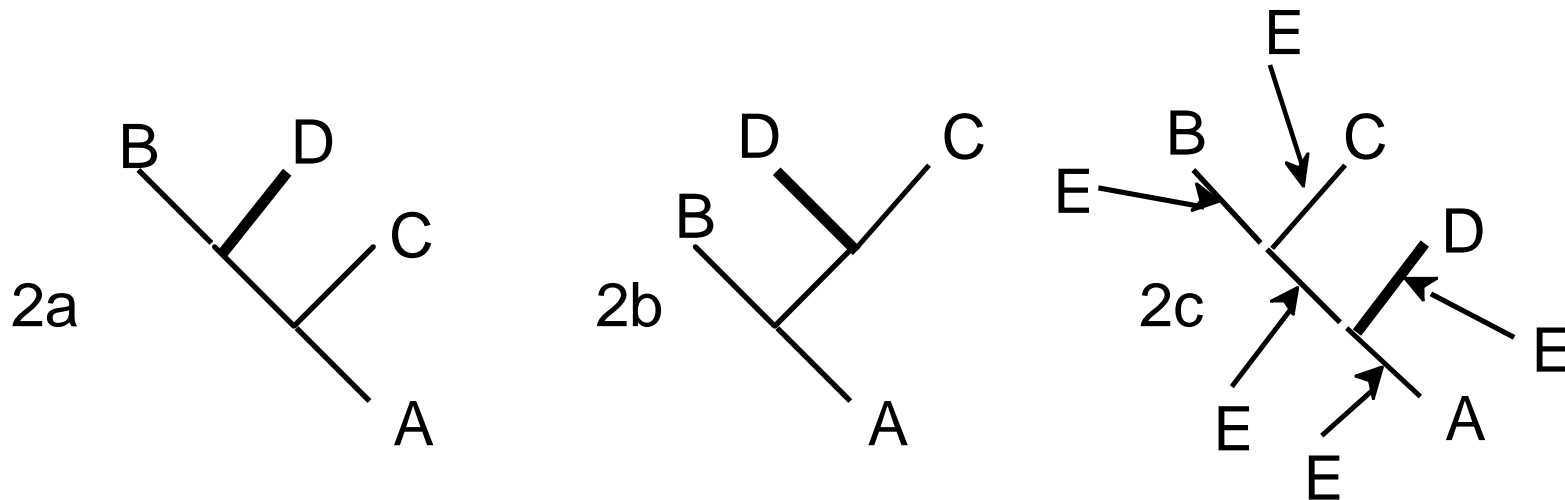
Finding optimal trees - exact solutions

- **Exact solutions can only be used for small numbers of taxa**
- **Exhaustive search examines all possible trees**
- **Typically used for problems with less than 10 taxa**

Finding optimal trees - exhaustive search



Add fourth taxon (D) in each of three possible positions -> three trees

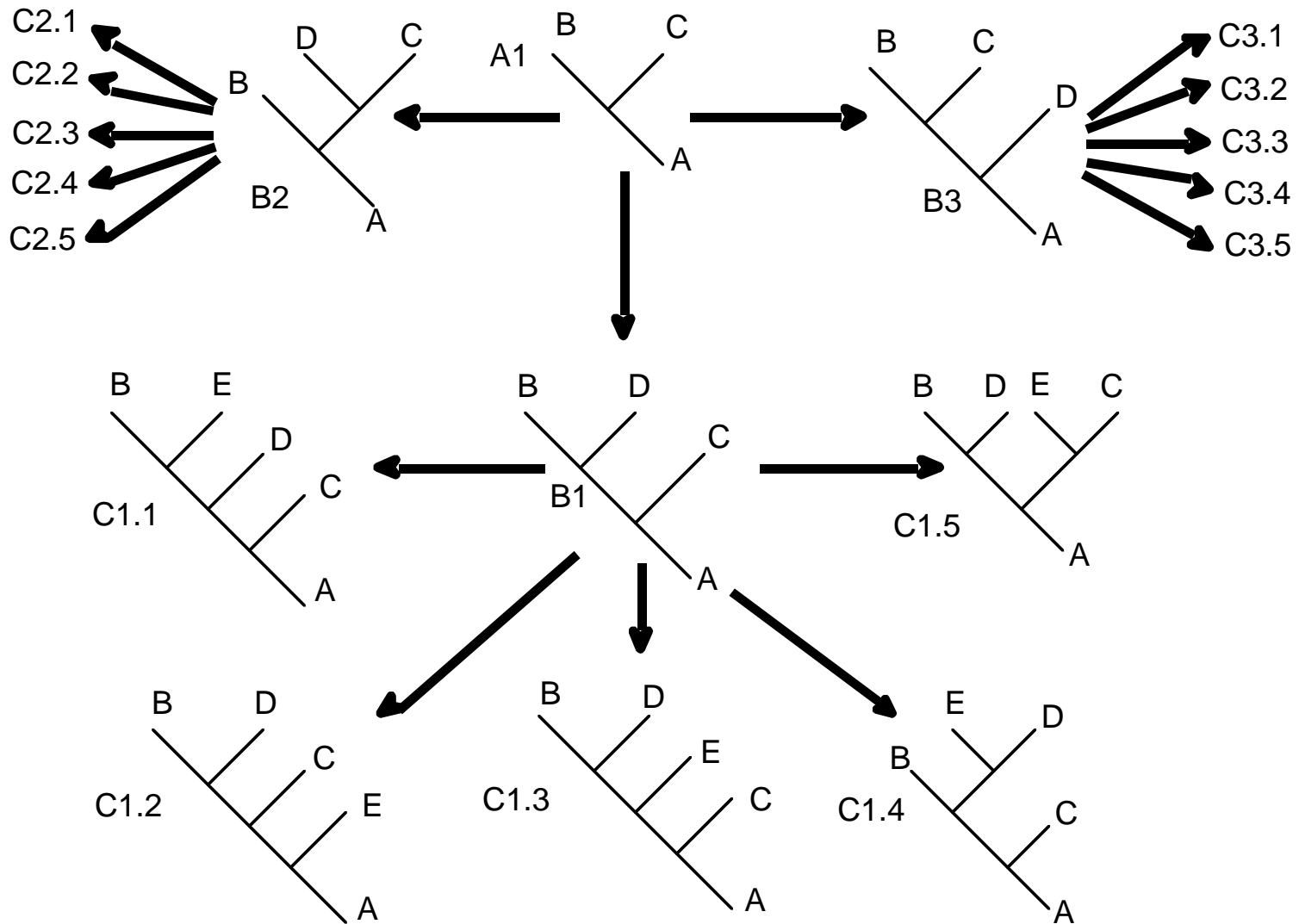


Add fifth taxon (E) in each of the five possible positions on each of the three trees -> 15 trees, and so on

Finding optimal trees - exact solutions

- **Branch and bound saves time by discarding families of trees during tree construction that cannot be shorter than the shortest tree found so far**
- **Can be enhanced by specifying an initial upper bound for tree length**
- **Typically used only for problems with less than 18 taxa**

Finding optimal trees - branch and bound



Finding optimal trees - heuristics

- The number of possible trees increases exponentially with the number of taxa making exhaustive searches impractical for many data sets (an NP complete problem)**
- Heuristic methods are used to search tree space for most parsimonious trees by building or selecting an initial tree and swapping branches to search for better ones**
- The trees found are not guaranteed to be the most parsimonious - they are best guesses**

Finding optimal trees - heuristics

- **Stepwise addition**

Asis - the order in the data matrix

Closest - starts with shortest 3-taxon tree adds taxa in order that produces the least increase in tree length (greedy heuristic)

Simple - the first taxon in the matrix is taken as a reference - taxa are added to it in the order of their decreasing similarity to the reference

Random - taxa are added in a random sequence, many different sequences can be used

- **Recommend random with as many (e.g. 10-100) addition sequences as practical**

Finding most parsimonious trees - heuristics

- **Branch Swapping:**

Nearest neighbor interchange (NNI)

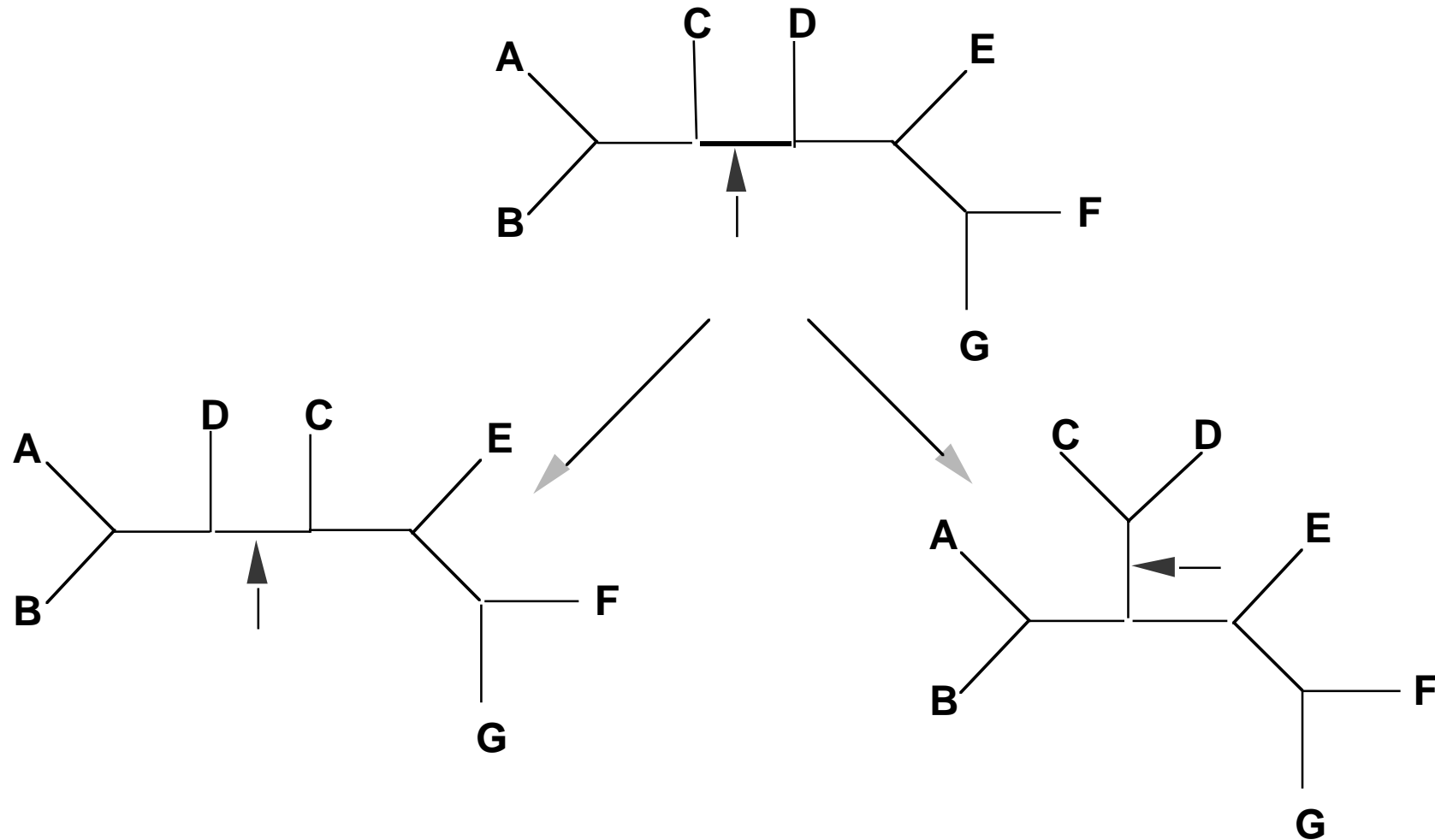
Subtree pruning and regrafting (SPR)

Tree bisection and reconnection (TBR)

Other methods

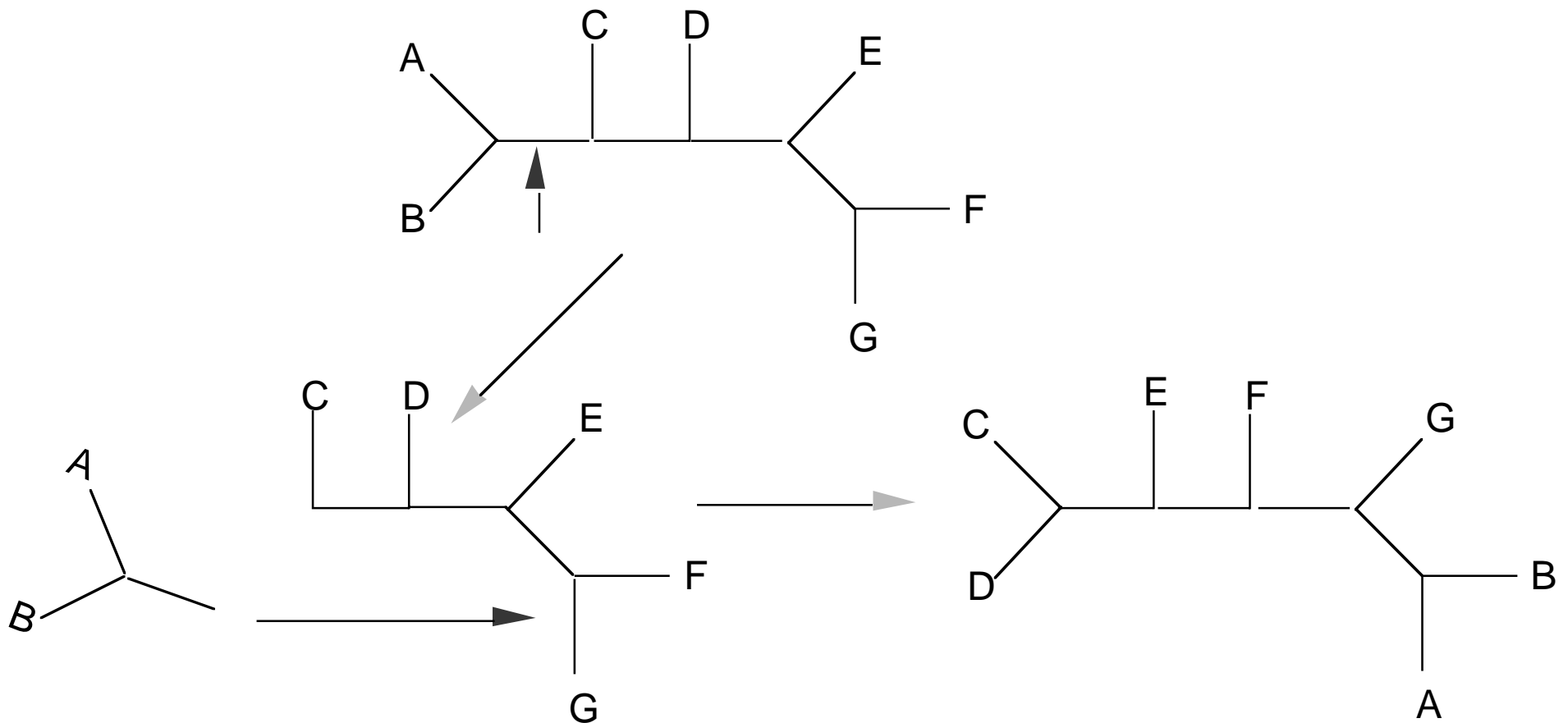
Finding optimal trees - heuristics

- **Nearest neighbor interchange (NNI)**



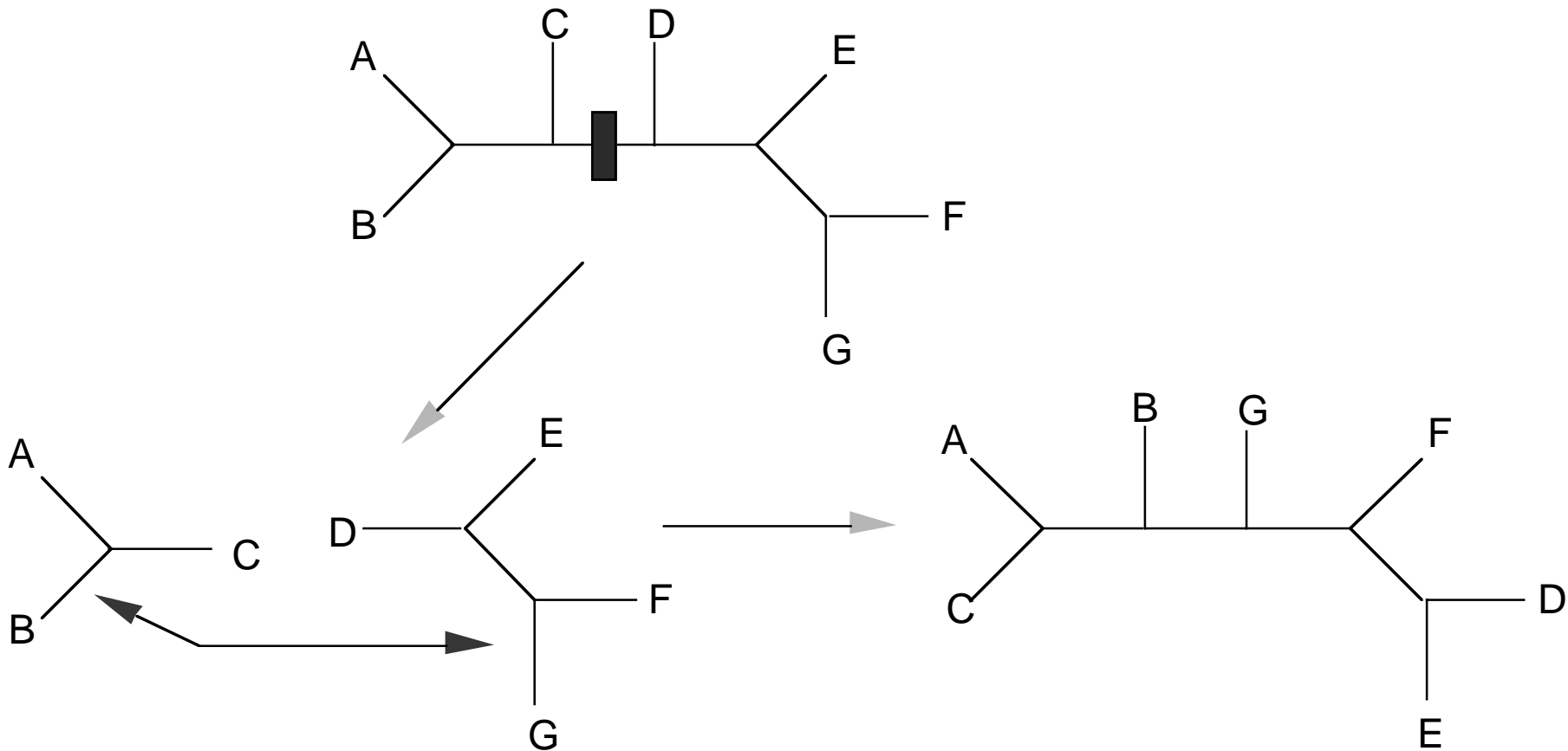
Finding optimal trees - heuristics

- Subtree pruning and regrafting (SPR)



Finding optimal trees - heuristics

- **Tree bisection and reconnection (TBR)**



Finding optimal trees - heuristics

- **Branch Swapping**

Nearest neighbor interchange (NNI)

Subtree pruning and regrafting (SPR)

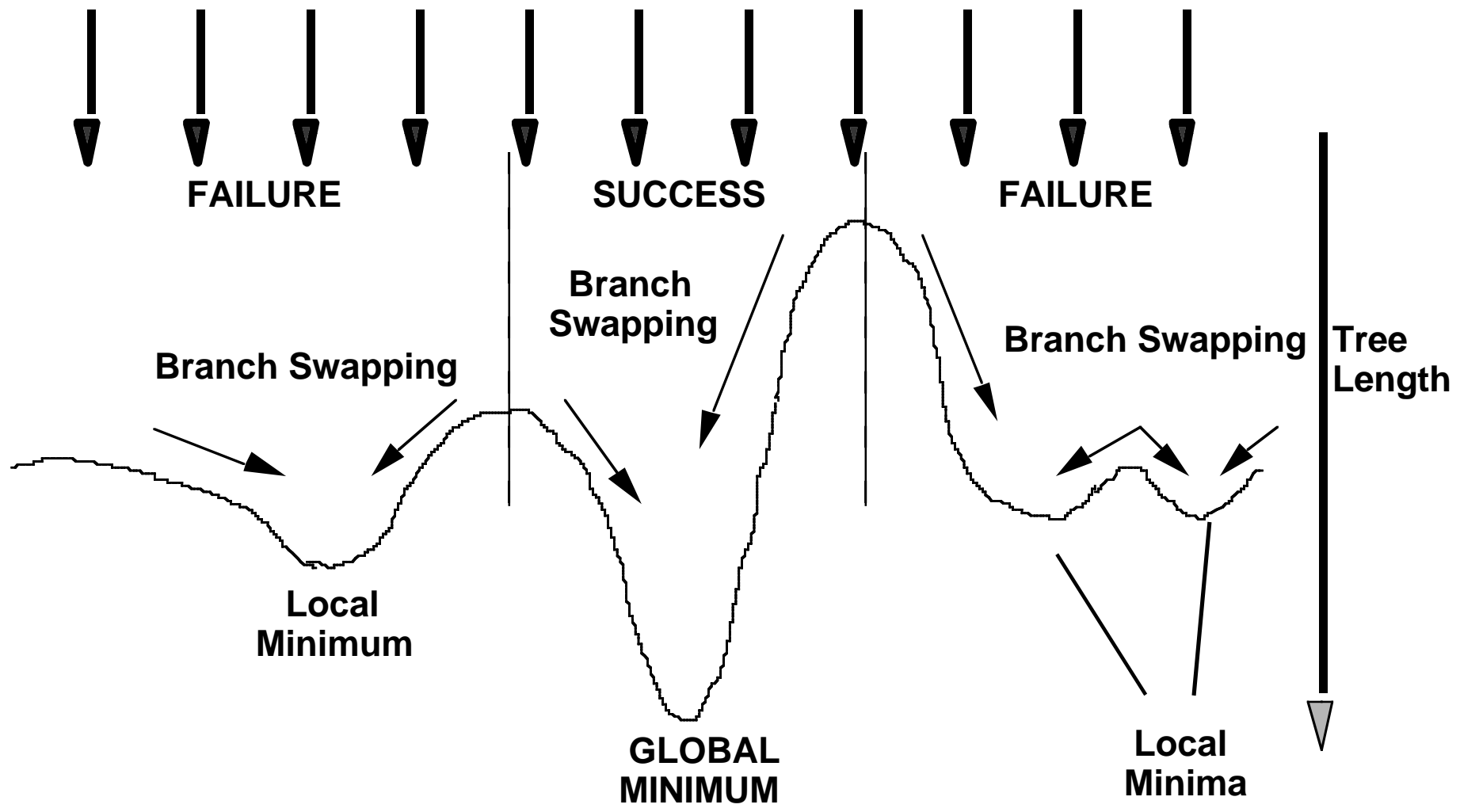
Tree bisection and reconnection (TBR)

- **The nature of heuristic searches means we cannot know which method will find the most parsimonious trees or all such trees**

- **However, TBR is the most extensive swapping routine and its use with multiple random addition sequences should work well**

Tree space may be populated by local minima and islands of optimal trees

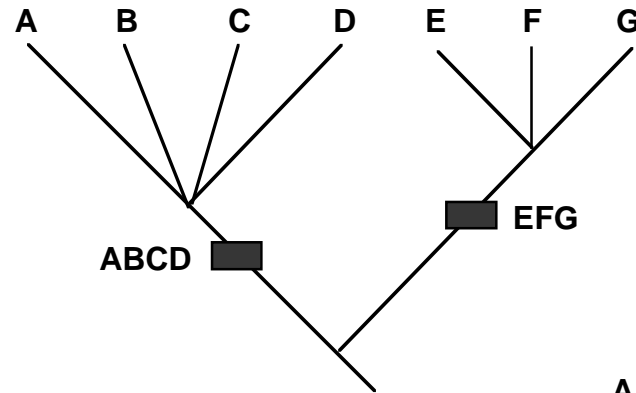
RANDOM ADDITION SEQUENCE REPLICATES



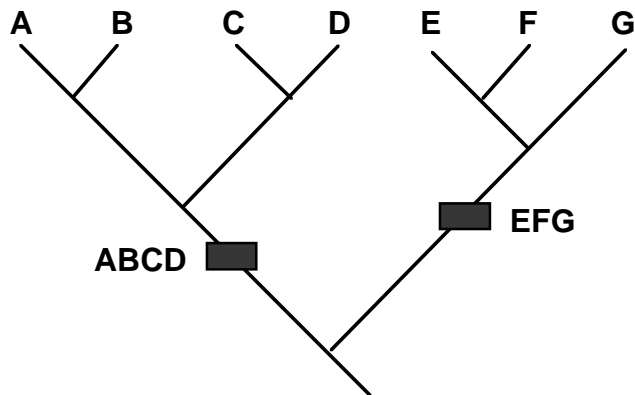
Searching with topological constraints

- **Topological constraints are user-defined phylogenetic hypotheses**
- **Can be used to find optimal trees that either:**
 - 1. include a specified clade or set of relationships**
 - 2. exclude a specified clade or set of relationships (reverse constraint)**

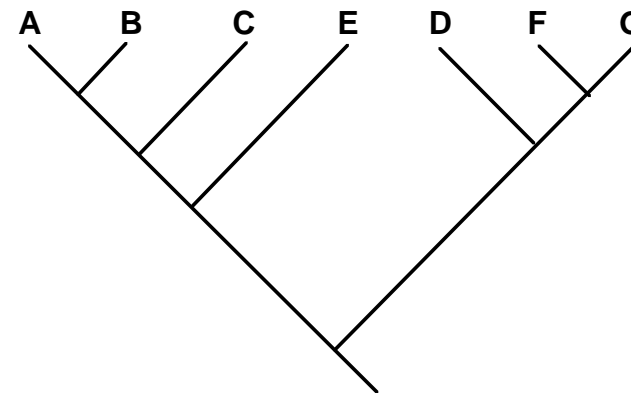
Searching with topological constraints



CONSTRAINT TREE
((A,B,C,D)(E,F,G))



Compatible with constraint tree
Incompatible with reverse constraint tree

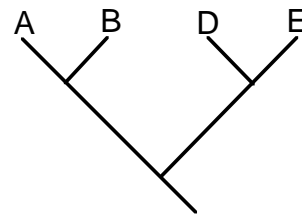


Compatible with reverse constraint tree
Incompatible with constraint tree

Searching with topological constraints

backbone constraints

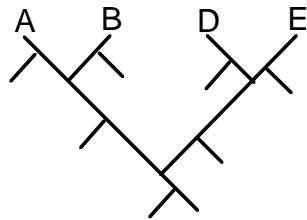
- **Backbone constraints specify relationships among a subset of the taxa**



BACKBONE CONSTRAINT

$((A,B)(D,E))$

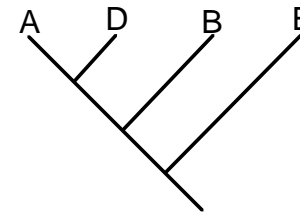
relationships of taxon C are not specified



/ possible positions of taxon C

Compatible with backbone constraint

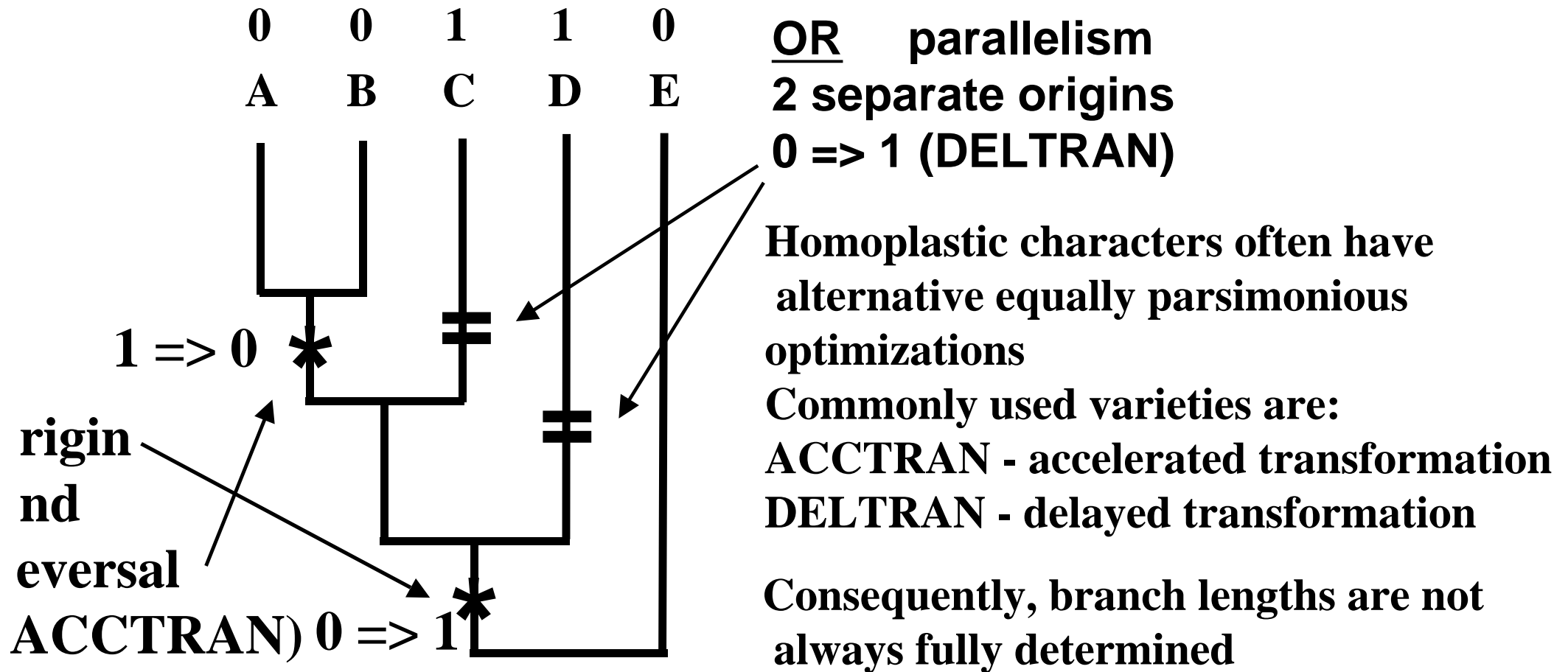
Incompatible with reverse constraint



Incompatible with backbone constraint

Compatible with reverse constraint

Parsimonious Character Optimization



PAUP reports minimum and maximum branch lengths

Multiple optimal trees

- **Many methods can yield multiple equally optimal trees**
- **We can further select among these trees with additional criteria, but**
- **Typically, relationships common to all the optimal trees are summarised with *consensus trees***

Consensus methods

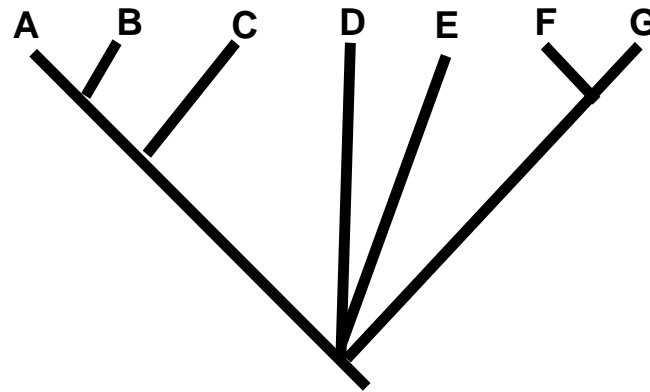
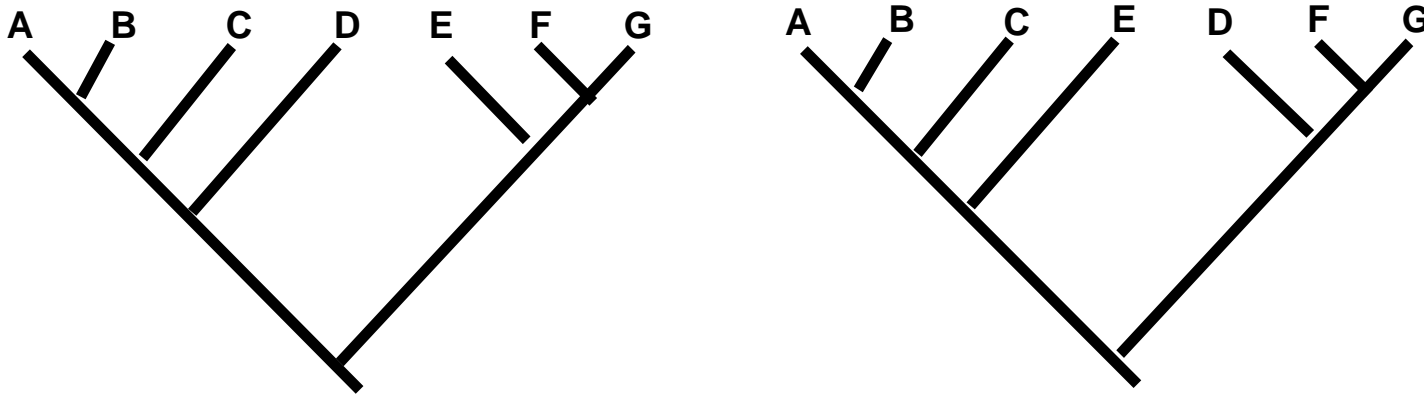
- **A consensus tree is a summary of the agreement among a set of fundamental trees**
- **There are many consensus methods that differ in:**
 - 1. the kind of agreement**
 - 2. the level of agreement**
- **Consensus methods can be used with multiple trees from a single analysis or from multiple analyses**

Strict consensus methods

- **Strict consensus methods require agreement across all the fundamental trees**
- **They show only those relationships that are unambiguously supported by the parsimonious interpretation of the data**
- **The commonest method (*strict component consensus*) focuses on clades/components/full splits**
- **This method produces a consensus tree that includes all and only those full splits found in all the fundamental trees**
- **Other relationships (those in which the fundamental trees disagree) are shown as unresolved polytomies**
- **Implemented in PAUP**

Strict consensus methods

TWO FUNDAMENTAL TREES



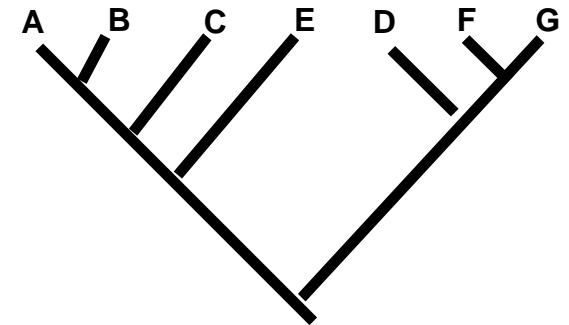
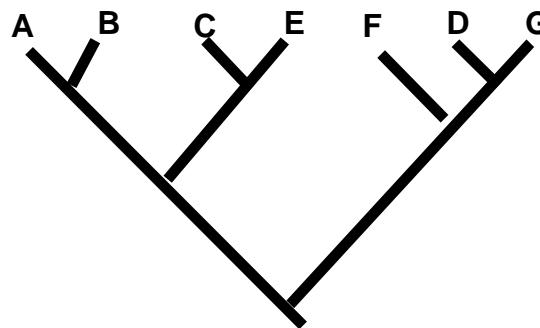
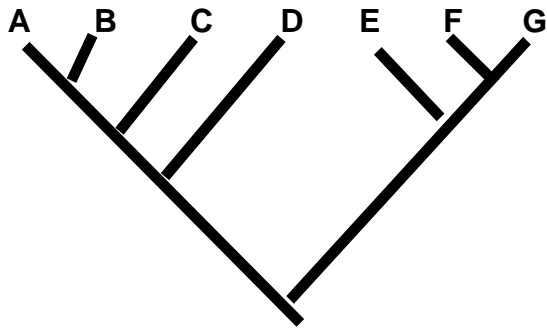
STRICT COMPONENT CONSENSUS TREE

Majority-rule consensus methods

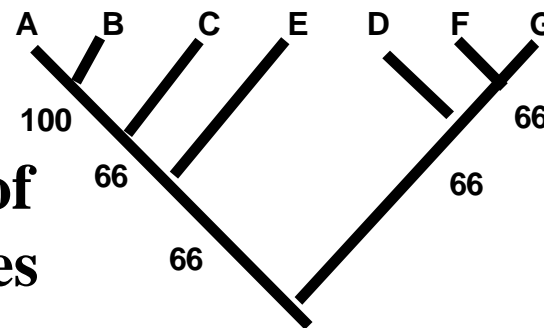
- **Majority-rule consensus methods require agreement across a majority of the fundamental trees**
- **May include relationships that are not supported by the most parsimonious interpretation of the data**
- **The commonest method focuses on clades/components/full splits**
- **This method produces a consensus tree that includes all and only those full splits found in a majority (>50%) of the fundamental trees**
- **Other relationships are shown as unresolved polytomies**
- **Of particular use in bootstrapping**
- **Implemented in PAUP**

Majority rule consensus

THREE FUNDAMENTAL TREES



Numbers indicate frequency of clades in the fundamental trees

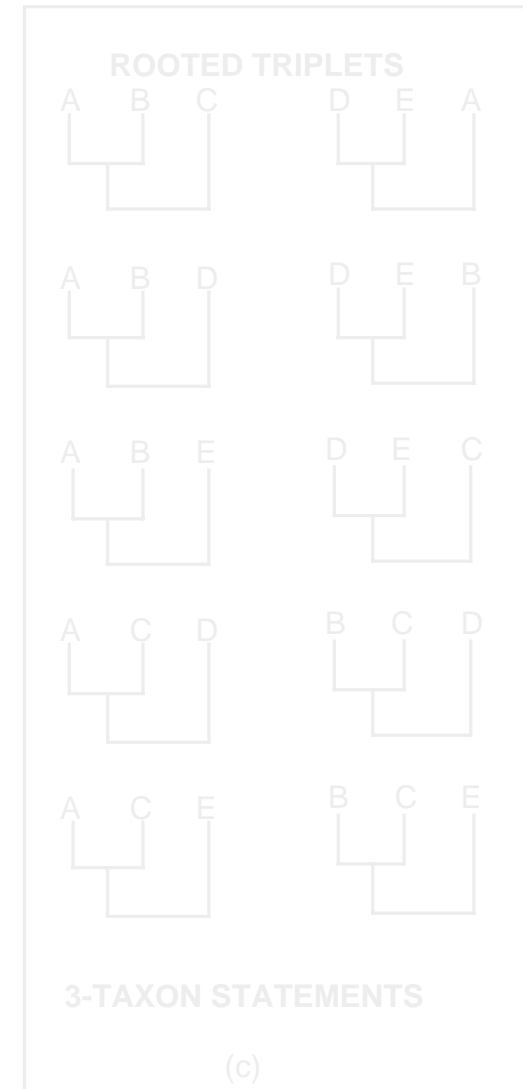
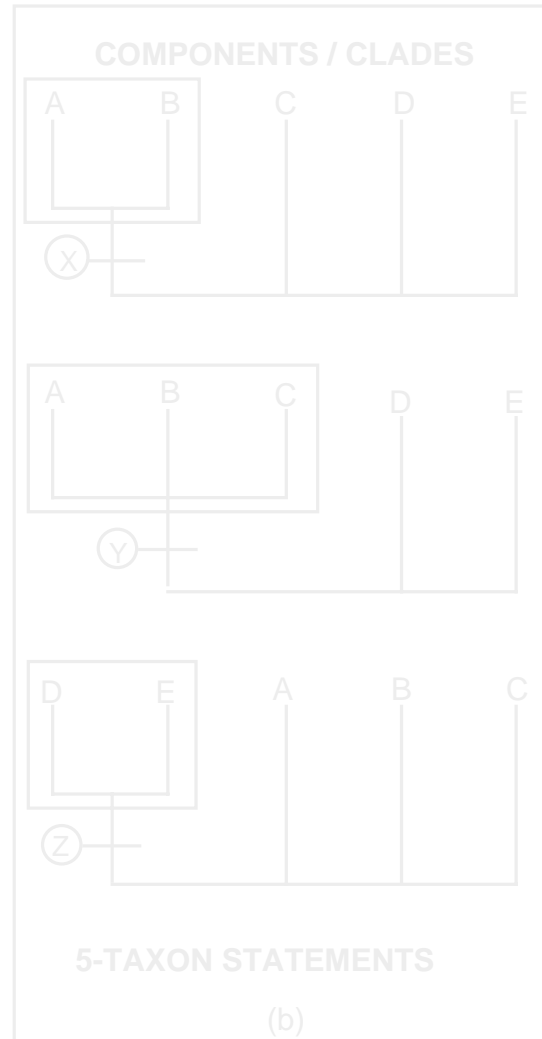
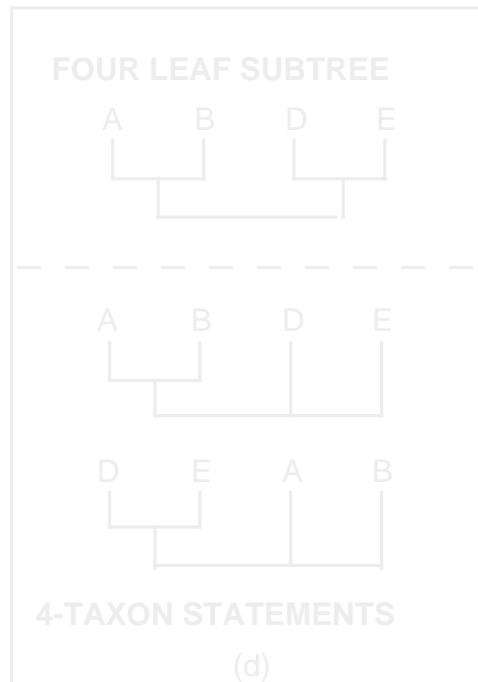
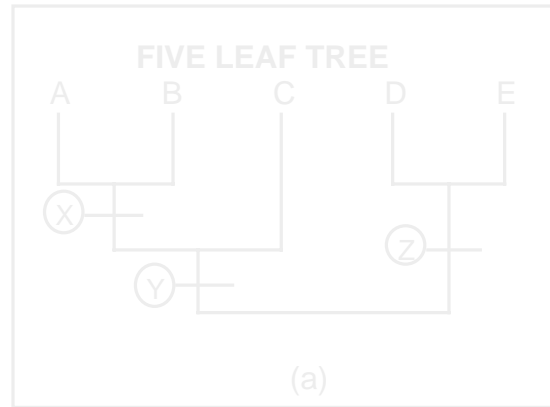


MAJORITY-RULE COMPONENT CONSENSUS TREE

Reduced consensus methods

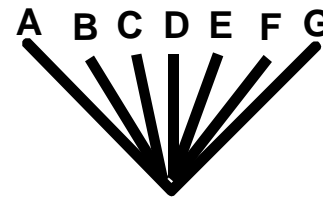
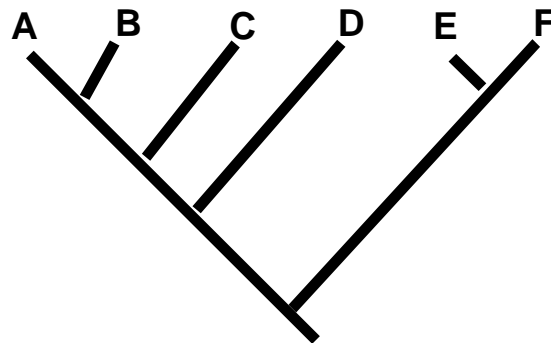
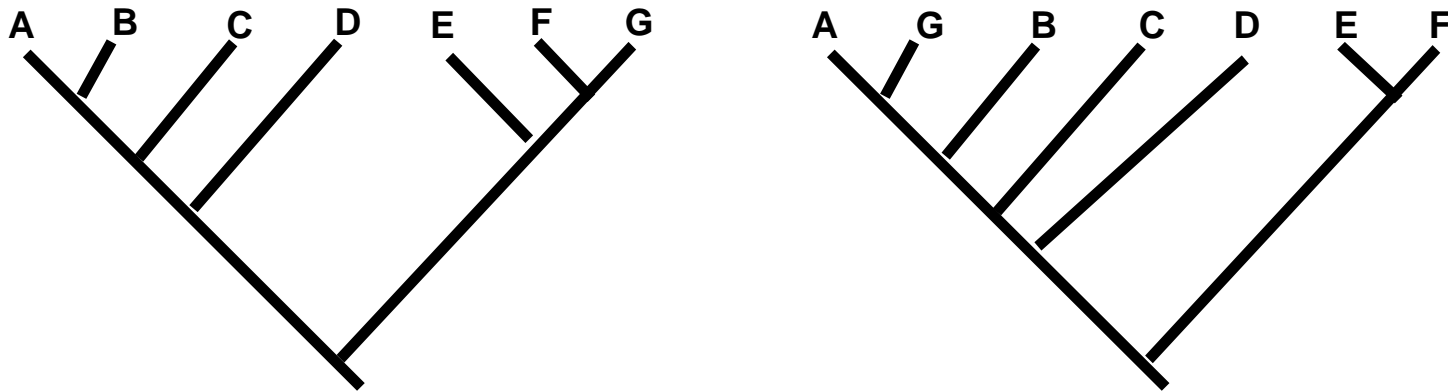
- **Focuses upon any relationships (not just full splits)**
- **Reduced consensus methods occur in strict and majority-rule varieties**
- **Other relationships are shown as unresolved polytomies**
- **May be more sensitive than methods focusing only on clades/components/full splits**
- **Strict reduced consensus methods are implemented in RadCon**

Types of Cladistic Relationships



Reduced consensus methods

TWO FUNDAMENTAL TREES



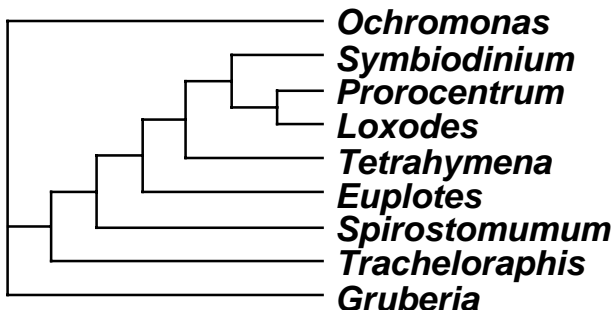
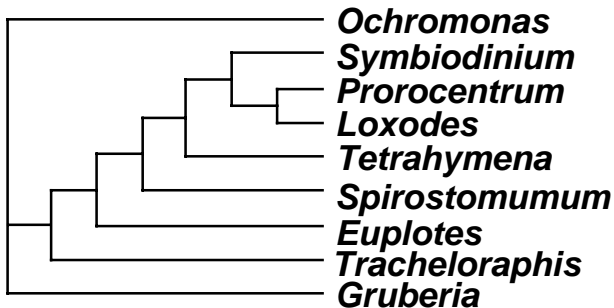
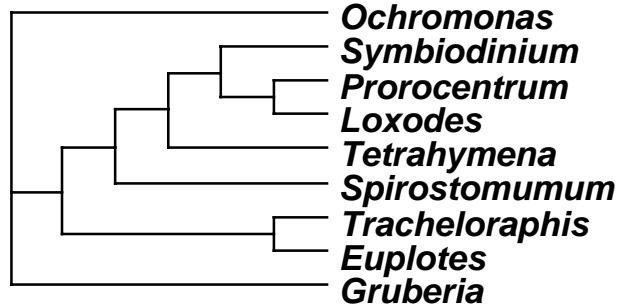
**Strict component consensus
completely unresolved**

STRICT REDUCED CONSENSUS TREE

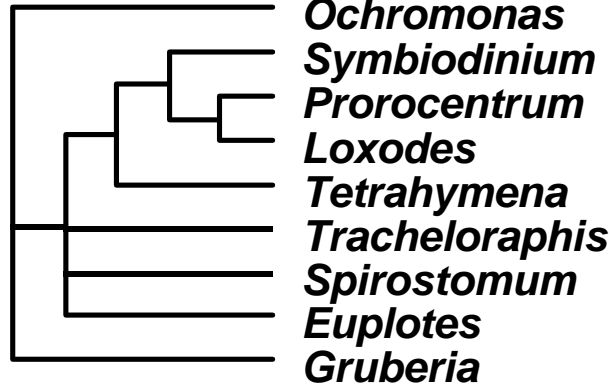
Taxon G is excluded

Consensus methods

Three fundamental trees

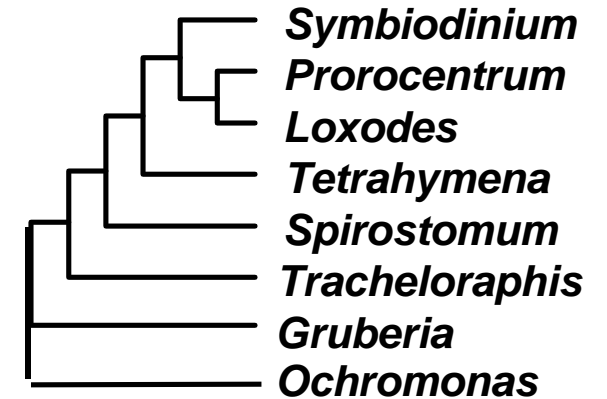


strict (component)

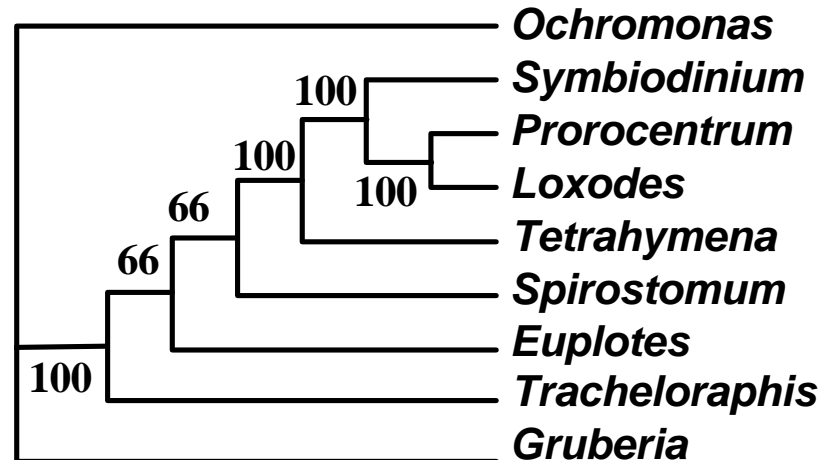


strict reduced cladistic

Euplotes excluded



majority-rule



Consensus methods

Use strict methods to identify those relationships unambiguously supported by parsimonious interpretation of the data

Use reduced methods where consensus trees are poorly resolved

Use majority-rule methods in bootstrapping

Avoid other methods which have ambiguous interpretations