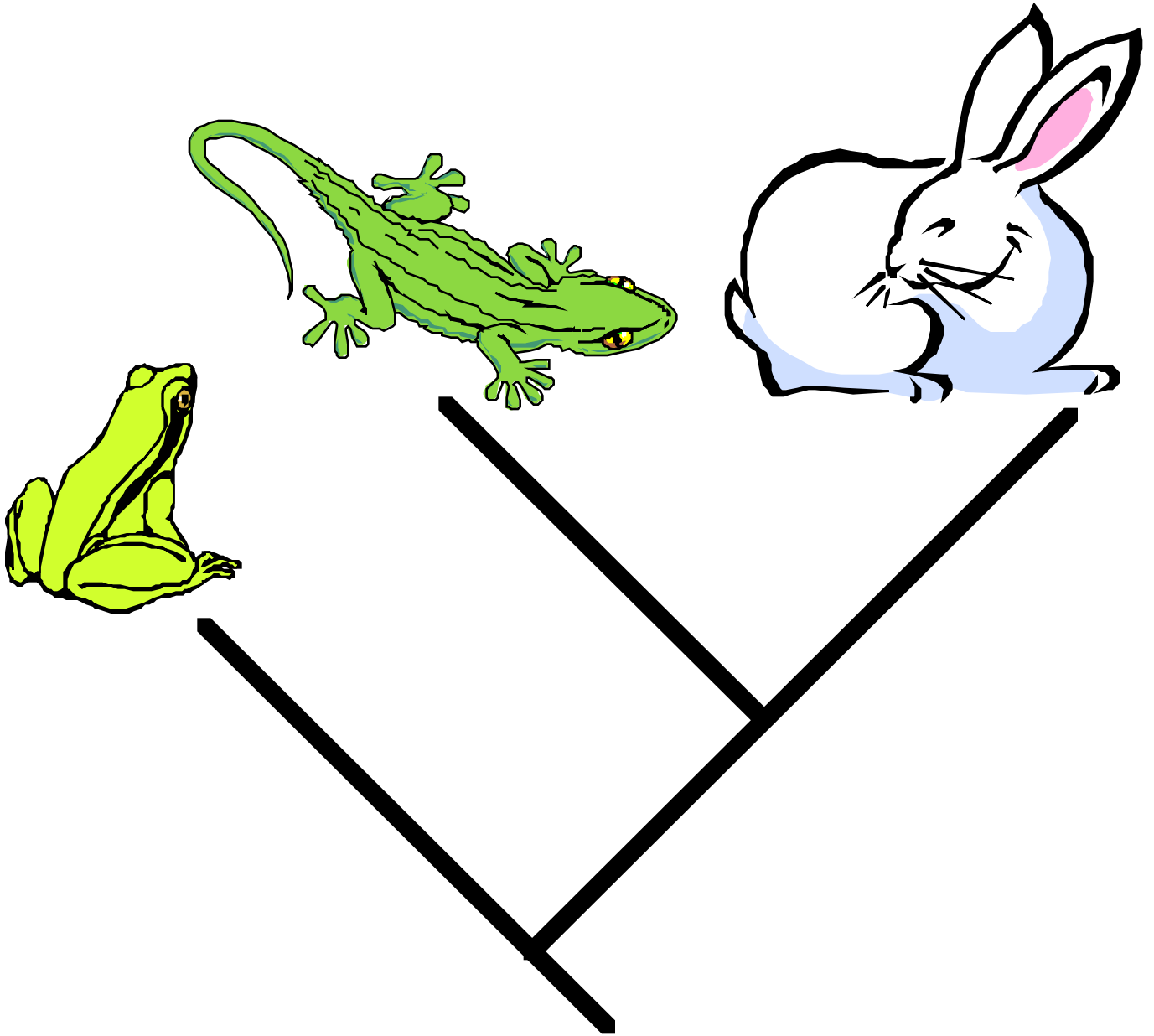


Basics of Cladistic Analysis



Diana Lipscomb
George Washington University
Washington D.C.
Copywrite (c) 1998

Preface

This guide is designed to acquaint students with the basic principles and methods of cladistic analysis. The first part briefly reviews basic cladistic methods and terminology. The remaining chapters describe how to diagnose cladograms, carry out character analysis, and deal with multiple trees. Each of these topics has worked examples.

I hope this guide makes using cladistic methods more accessible for you and your students. Report any errors or omissions you find to me and if you copy this guide for others, please include this page so that they too can contact me.

Diana Lipscomb

Weintraub Program in Systematics &
Department of Biological Sciences
George Washington University
Washington D.C. 20052 USA
e-mail: BIODL@GWU.EDU

Introduction to Systematics

All of the many different kinds of organisms on Earth are the result of evolution. If the evolutionary history, or **phylogeny**, of an organism is traced back it connects through shared ancestors to lineages of other organisms. That all of life is connected in an immense **phylogenetic tree** is one of the most significant discoveries of the past 150 years. The field of biology that reconstructs this tree and uncovers the pattern of events that led to the distribution and diversity of life is called **systematics**. Systematics, then, is no less than understanding the history of all life.

In addition to the obvious intellectual importance of this field, systematics forms the basis of all other fields of comparative biology:

- *Systematics provides the framework, or **classification**, by which other biologists communicate information about organisms*
- *Systematics and its phylogenetic trees provide the basis of evolutionary interpretation*
- *The phylogenetic tree and corresponding classification predicts properties of newly discovered or poorly known organisms*

THE SYSTEMATIC PROCESS

The systematic process consists of five interdependent but distinct steps:

1. The taxa to be classified are chosen.
2. The characteristics to provide evidence for relationship are chosen.
3. The characteristics are analyzed to reconstruct the relationship among the taxa (usually in the form of a tree).
4. The tree is translated into a formal classification system so that it can be communicated and used by other scientists.

5. The tree is used to test various hypotheses about the process of evolution in the group.

The primary purpose of this booklet is to guide you through step 3. So, for now, we will skip choosing taxa and characters (steps 1 and 2) and go straight to cladistic analysis.

Cladistics or Phylogenetic Systematics

Given that closely related species share a common ancestor and often resemble each other, it might seem that the best way to uncover the evolutionary relationships would be with overall similarity. In other words, out of a group of species, if two are most similar, can we reasonably hypothesize that they are closest relatives? Perhaps surprisingly, the answer is **no**. Overall similarity may be misleading because there are actually two reasons why organisms have similar characteristics and only one of them is due to evolutionary relatedness. When two species have a similar characteristic because it was inherited by both from a common ancestor, it is called a **homologous features** (or **homology**). For example, the even-toed foot of the deer, camels, cattle, pigs, and hippopotamus is a homologous similarity because all inherited the feature from their common paleodont ancestor.

On the other hand, when unrelated species adopt a similar way of life, their body parts may take on similar functions and end up resembling one another due to convergent evolution. When two species have a similar characteristic because of convergent evolution, the feature is called an **analogous features** (or **homoplasy**). The paddle-like front limb and streamlined bodies of the aquatic animals shown in the figure for convergent evolution are examples of analogous features.

Only homologous similarity is evidence that two species are evolutionarily related. If two animals share the highest number of homologies, can we reasonably assume they are closest relatives? The answer is still **no** - a homology may be recently derived or an ancient retained feature; only shared recent homologies (called synapomorphies) are evidence that two organisms are closely related. An example will make this point clear: The hand of the first vertebrates to live on land had five digits (fingers). Many living terrestrial vertebrates (such as humans, turtles, crocodiles and frogs) also have five digits because they inherited them from this common ancestor. This feature is then homologous in all of these species. In

contrast, horses, zebras and donkeys have just a single digit with a hoof. Clearly, humans are more closely related to horses, zebras and donkeys, even though they have a homology in common with turtles, crocodiles and frogs. The key point is that the five digit condition is the primitive state for the number of digits. It was modified and reduced to just one digit in the common ancestor of horses, donkeys and zebras. The modified derived state does tell us that horses, zebras and donkeys share a very recent common ancestor, but the primitive form is not evidence that species are particularly closely related.

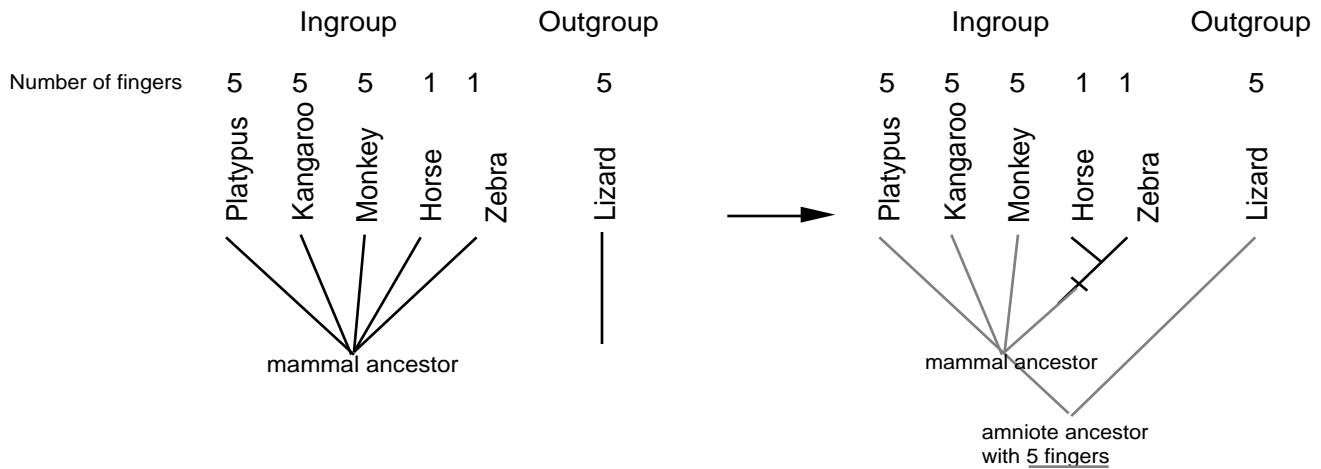
In an important work (first published in English in 1966) by the German entomologist Willi Hennig, it was argued that only shared derived characters could possibly give us information about phylogeny. The method that groups organisms that share derived characters is called **cladistics** or **phylogenetic systematics**. Taxa that share many derived characters are grouped more closely together than those that do not. The relationships are shown in a branching hierarchical tree called a cladogram. The cladogram is constructed such that the number of changes from one character state to the next are minimized. The principle behind this is the rule of **parsimony** - any hypothesis that requires fewer assumptions is a more defensible hypothesis.

DETERMINING PRIMITIVE (PLESIOMORPHIC) AND DERIVED (APOMORPHIC) CHARACTERS

The first step in basic cladistic analysis is to determine which character states are primitive and which are derived. Many methods for doing this were proposed by Hennig (1966) and others, but the **outgroup comparison** method is the primary one in use today. In outgroup comparison, if a taxon that is not a member of the group of organisms being classified has a character state that is the same as some of the organisms in the group, then that character state can be considered to be plesiomorphic. The outside taxon is called the **outgroup** and the organisms being classified are the **ingroup**.

Two arguments can be made to justify using this method: one based on what we believe about evolutionary process and the other on logic:

1. The only way a homologous feature could be present in both an ingroup and an outgroup, would be for it to have been inherited by both from an ancestor older than the ancestor of just the ingroup:



2. Consider the following example in which a character has states **a** and **b**. There are only two possibilities:

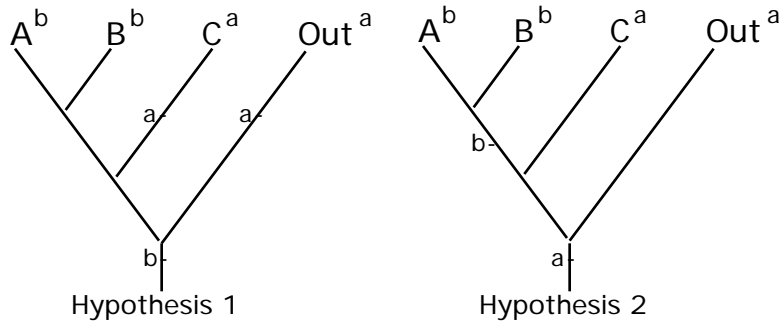
1. **b** is plesiomorphic and **a** is apomorphic

b --> **a**

2. **a** is plesiomorphic and **b** is apomorphic

a --> **b**

If state **a** is also found in a taxon outside the group being studied, the first hypothesis will force us to make more assumptions than the second (it is less parsimonious):



In hypothesis 1, state **b** is the plesiomorphic state and is placed at the base of the tree. If state **b** only evolved once, it must be assumed that character state **a** evolved two separate times - once in taxon C and once in the Outgroup. In hypothesis 2, state **a** is the plesiomorphic state and is placed at the base of the tree. In this hypothesis, both character states **a** and **b** each evolve only once. Therefore, hypothesis 2 is more parsimonious and is a more defensible hypothesis. This example illustrates why outgroup analysis gives the most parsimonious, and therefore logical, hypothesis of which state is plesiomorphic.

If the character has only two states, then the task of distinguishing primitive and derived character states is fairly simple: The state which is in the outgroup is primitive and the one found only in the ingroup is derived.

It is common practice to designate the primitive states as 0 (zero) and the derived states as 1 (one). If you are going to calculate trees by hand, this will certainly make your calculations easier. On the other hand, if you are using a computer program to calculate a tree, it isn't necessary to designate the plesiomorphic state as 0 (zero):

	Characters											Characters									
	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
Outgroup	a	a	a	a	b	a	a	a	a	a	Outgroup	0	0	0	0	0	0	0	0	0	0
Alpha	a	a	a	b	a	b	a	a	a	a	Alpha	0	0	0	1	1	1	0	0	0	0
Beta	a	a	a	b	b	b	a	a	a	a	Beta	0	0	0	1	0	1	0	0	0	0
Gamma	a	a	a	a	b	a	b	a	b	a	Gamma	0	0	0	0	0	0	1	0	1	0
Delta	b	b	b	a	b	a	b	b	b	a	Delta	1	1	1	0	0	0	1	1	1	0
Epsilon	b	b	b	a	b	a	b	a	b	b	Epsilon	1	1	1	0	0	0	1	0	1	1
Zeta	b	b	a	a	b	a	b	a	b	b	Zeta	1	1	0	0	0	0	1	0	1	1
Theta	b	a	a	a	b	a	b	a	b	b	Theta	1	0	0	0	0	0	1	0	1	1

CONSTRUCTING A CLADOGRAM

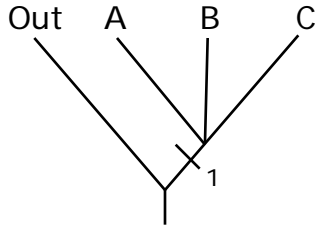
There are two ways to group taxa together based on shared apomorphies. The first, called **Hennig Argumentation**, is essentially the method as described by Hennig in his 1966 book. The other method, called the **Wagner Method**, uses an algorithm to search for the tree and was developed by Kluge and Farris (1969) and Farris (1970).

Hennig Argumentation

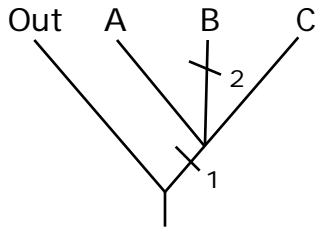
Hennig argumentation considers the information provided by each character one at a time. This is easiest to understand with a small data set:

	Characters				
	1	2	3	4	5
Outgroup	0	0	0	0	0
A	1	0	0	0	1
B	1	1	0	1	0
C	1	0	1	1	0

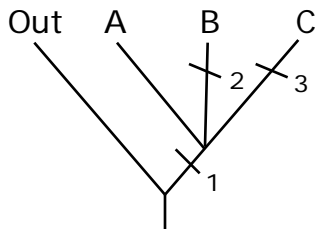
1. The information in character 1 unites taxa A, B, and C because they share the apomorphic state. The tree that shows this relationship is:



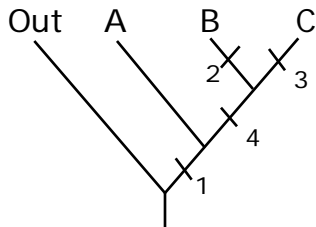
2. Character 2 - the derived state is found only in taxon B. It is an **autapomorphy** of that taxon and provides no information about the relationships among the taxa:



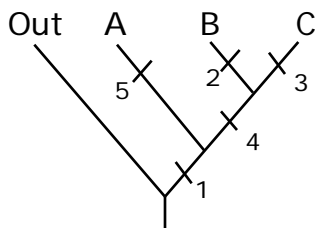
3. Character 3 - the derived state is an autapomorphy for taxon C:



4. Character 4 - the derived state is a synapomorphy that unites taxa B and C:



5. Character 5 - the derived state is an autapomorphy for taxon A:



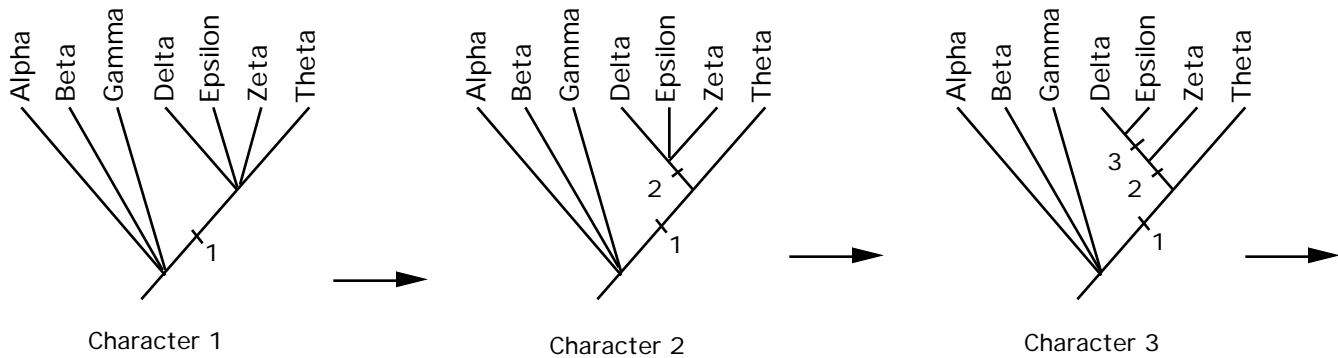
The cladogram is finished - All characters have been considered and the relationships of the taxa are resolved.

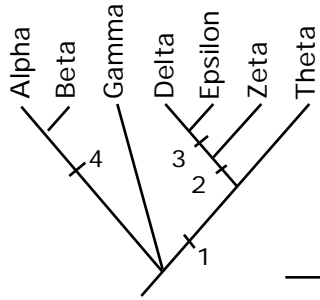
Real sets of character data are rarely so simple. A more realistic situation would be one where the characters conflict about the relationships among the taxa. When it does, choose the tree that requires the fewest assumptions about character states changing.

To examine this, construct a cladogram using Hennig argumentation from the more complicated data set:

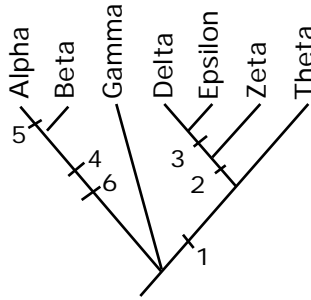
	Characters									
	1	2	3	4	5	6	7	8	9	10
Outgroup	0	0	0	0	0	0	0	0	0	0
Alpha	0	0	0	1	1	1	0	0	0	0
Beta	0	0	0	1	0	1	0	0	0	0
Gamma	0	0	0	0	0	0	1	0	1	0
Delta	1	1	1	0	0	0	1	1	1	0
Epsilon	1	1	1	0	0	0	1	0	1	1
Zeta	1	1	0	0	0	0	1	0	1	1
Theta	1	0	0	0	0	0	1	0	1	1

1. Grouping taxa is straightforward when characters 1-9 are used:

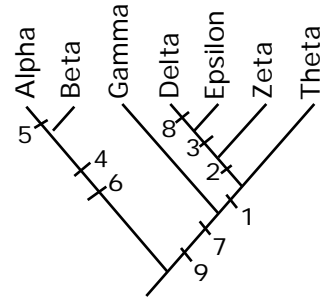




Character 4

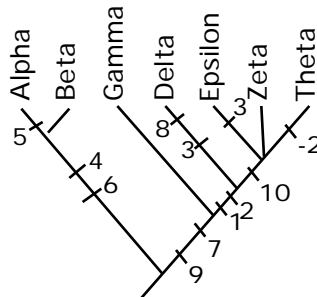


Character 5 is an autapomorphy (found only in Alpha), and character 6 has the same distribution as character 4



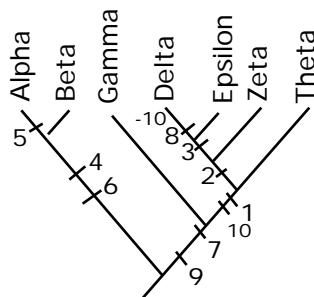
Character 8 is an autapomorphy (found only in Delta), and character 7 and 9 have the same distribution

2. Character 10 conflicts with characters 2 and 3 - it suggests Epsilon, Zeta and Theta are sister taxa, whereas 2 puts Epsilon and Zeta closer to Delta than to Theta:



If the tree that conforms to character 10 tree is accepted, we must assume that character 2 secondarily reverses to state 0 in taxon Theta and that character 3 is convergent in Delta and Epsilon. In other words it requires us to assume 2 homoplasious steps (one reversal and one convergence).

On the other hand, if the tree that characters 2 and 3 support is accepted, we must assume that character 10 reverses to state 0 in taxon Delta:



Because this latter tree requires us to make the fewest assumptions, it is the most parsimonious and therefore is the preferred cladogram.

WAGNER TREES

A second way to construct a cladogram is to connect taxa together one at a time until all the taxa have been added. When added, each taxon is joined to the tree to minimize the number of character state changes.

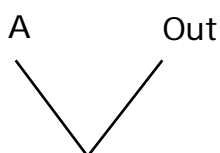
Consider again the small data set:

	Characters				
	1	2	3	4	5
Outgroup	0	0	0	0	0
A	1	0	0	0	0
B	1	1	0	1	0
C	1	0	1	1	1

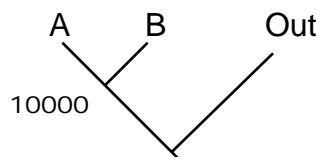
1. Find the organism with the lowest number of derived character states and connect it to the outgroup.

	Characters					#advanced steps
	1	2	3	4	5	
Outgroup	0	0	0	0	0	0
A	1	0	0	0	0	1
B	1	1	0	1	0	3
C	1	0	1	1	1	4

Organism A has the lowest number of advanced steps:

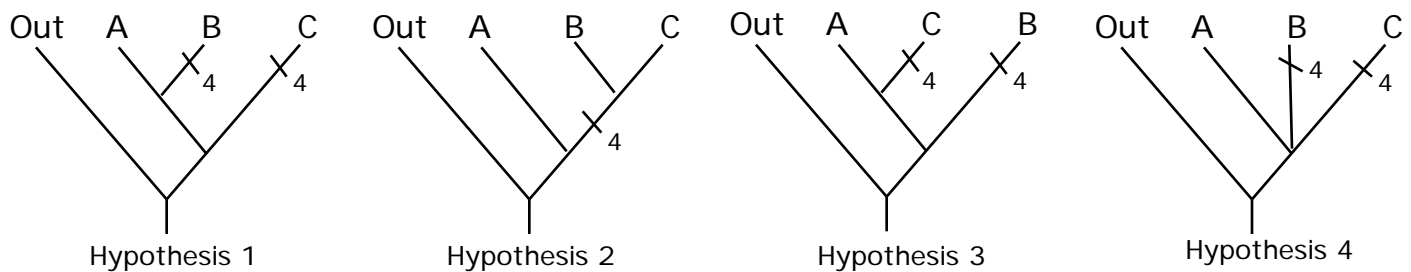


2. Now find the organism with the next lowest number of derived character states. Write its name beside the first organisms' name and connect it to the line that joins the outgroup and the first organism. At the point where the two lines intersect, list the most advanced state present in both of the two organisms. For the above character set, the second organism is B:



Organisms A and B both have the derived state for character 1 but one or both of them have state 0 for the remaining characters. At the point where the two lines intersect, 1000 is given (this is called optimization, and we will return to it later).

3. Find the organism with the next lowest number of derived character states and connect it to the point which requires the fewest number of evolutionary steps (character state changes) to derive the organism. In this example, organism C is added next. There are several different places it can be attached:



Hypotheses 1, 3, and 4 all require us to assume that character 4 state 1 evolved two times. Hypothesis 2 does not require us to make that assumption. Therefore the most parsimonious placement of organism C is as in Hypothesis 2. The analysis is complete.

Wagner Formula

The number of steps required to attach a taxon to any other can be summarized by the formula:

$$d(A, B) = \sum |X(A_i) - X(B_i)|$$

Where,

d = the number of character state changes between taxa A and B

= sum of all differences between

X(A,*i*) = the state of a character for taxon A
and

X(B,*i*) = the state of a character for taxon B

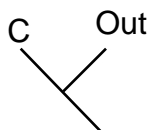
A simple way to see how this can be calculated is by working through an example

using the data set:

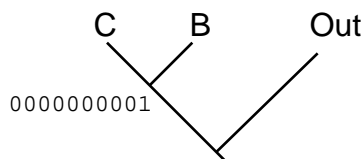
	Characters										# advanced steps
	1	2	3	4	5	6	7	8	9	10	
Out	0	0	0	0	0	0	0	0	0	0	
A	1	0	0	0	1	1	0	0	0	1	4
B	1	0	0	0	1	0	0	0	0	1	3
C	0	0	0	0	0	0	0	0	1	1	2
D	0	1	1	0	0	0	1	1	0	1	5
E	0	1	1	1	0	0	0	1	0	1	5

1. Find the organism with the lowest number of derived character states and connect it to the outgroup.

C has the fewest number of derived steps. It is linked to the outgroup first:



2. B has the next fewest number of advanced states. It is joined to C. At the point where the two lines intersect, the advanced states present in both are listed. For example, character 1 is at state 0 in C but is state 1 for B. 0 is the most advanced state that they share so it is listed next to the point where the lines leading to C and B intersect. On the other hand, both B and C have state 1 for character 10. Therefore, a 1 is given for character 10 at the point where the two lines intersect. Listing the most plesiomorphic states shared by the taxa at the node that they share, creates a path from the node to the taxon that requires the fewest overall numbers of character state changes (I.e., is most parsimonious)



3. A has the next lowest number of advanced character states. It must be connected such that the fewest number of steps are required to reach it (fewest number of character state changes). Here we begin to apply the Wagner formula.

If A were attached to the line leading to C:

A 1000110001
C 0000000011
differences: 1 11 1 (4 steps would have to be made)

To attach **A** to the line leading to **C** would require the assumptions that characters 1, 5, 6, and 9 all evolved advanced states.

If **A** were attached to the line leading to **B**:

A 1000110001
B 1000100001
differences: 1 (1 step would have to be made)

To attach **A** to the line leading to **B** would require the assumptions that character 6 evolved an advanced state.

If **A** were attached to the line leading to the point where the lines leading to **B** and **C** intersect:

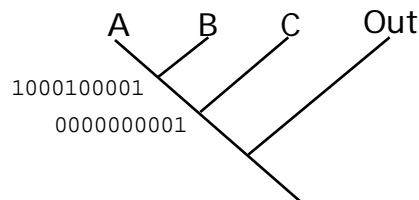
A 1000110001
BC 0000000001
differences: 1 11 (3 steps would have to be made)

To attach **A** to the line leading to **BC** would require the assumptions that characters 1, 5, and 6 all evolved advanced states.

If **A** were attached to the line leading to the Outgroup:

A 1000110001
O 0000000000
differences: 1 11 1 (4 steps would have to be made)

To attach **A** to the line leading to the outgroup would require the assumptions that characters 1, 5, 6, and 10 all evolved advanced states. **Since it would take the fewest evolutionary steps to attach A to B, that is what is done:**



4. Either **D** or **E** can be attached next. For this example, **D** will be attached first, then **E**.

If **D** were attached to the line leading to **A**:

```
D    0110001101
A    1000110001
differences: 111 1111    (7 steps would have to be made)
```

To attach **D** to the line leading to **A** would require the assumptions that characters 1,2,3,5,6,7, and 8 evolved advanced states.

If **D** were attached to the line leading to **B**:

```
D    0110001101
B    1000100001
differences: 111 1 11    (6 steps would have to be made)
```

To attach **D** to the line leading to **B** would require the assumptions that characters 1,2,3,5,7, and 8 evolved advanced states.

If **D** were attached to the line leading to **C**:

```
D    0110001101
C    0000000011
differences: 11 111    (5 steps would have to be made)
```

To attach **D** to the line leading to **C** would require the assumptions that characters 2, 3, 7, 8, and 9 evolved advanced states.

If **D** were attached to the line leading to the point where the lines leading to **B** and **A** intersect:

```
D    0110001101
AB   1000100001
differences: 111 1 11    (6 steps would have to be made)
```

To attach **D** to the line leading to **AB** would require the assumptions that characters 1, 2, 3, 5, 7 and 8 all evolved advanced states.

If **D** were attached to the line leading to the point where the lines leading to **B**, **A** and **C** intersect:

```
D    0110001101
ABC  0000000001
differences: 11 11    (4 steps would have to be made)
```

To attach **D** to the line leading to **ABC** would require the assumptions that characters 2, 3, 7 and 8 all evolved advanced states.

If **D** were attached to the line leading to the Outgroup:

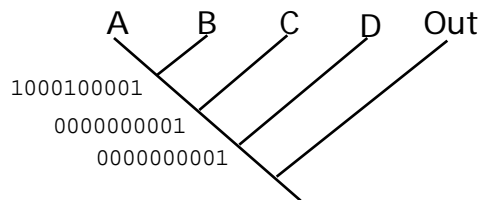
```

D      0110001101
O      0000000000
differences: 11 11 1 (5 steps would have to be made)

```

To attach **D** to the line leading to the outgroup would require the assumptions that characters 1, 5, 6, and 10 all evolved advanced states.

Since it would take the fewest evolutionary steps to attach **D** to the line leading to **ABC**, that is what is done:



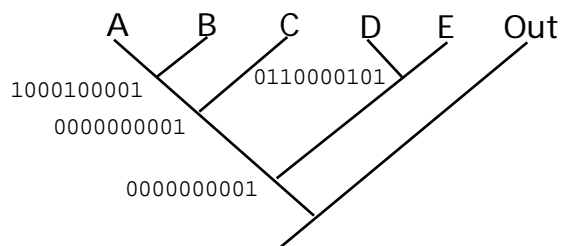
5. **E** is attached last:

```

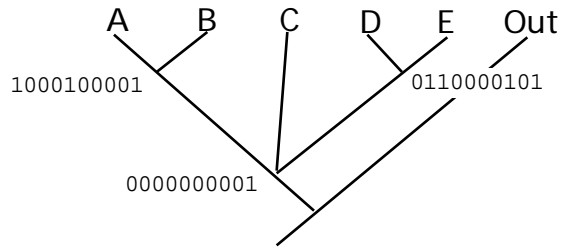
E --> A      = 7 steps
E --> B      = 6 steps
E --> C      = 5 steps
E --> D      = 2 steps
E --> AB     = 6 steps
E --> ABC    = 4 steps
E --> ABCD   = 4 steps
E --> Out    = 5 steps

```

E is connected to **D** and the cladogram is complete:



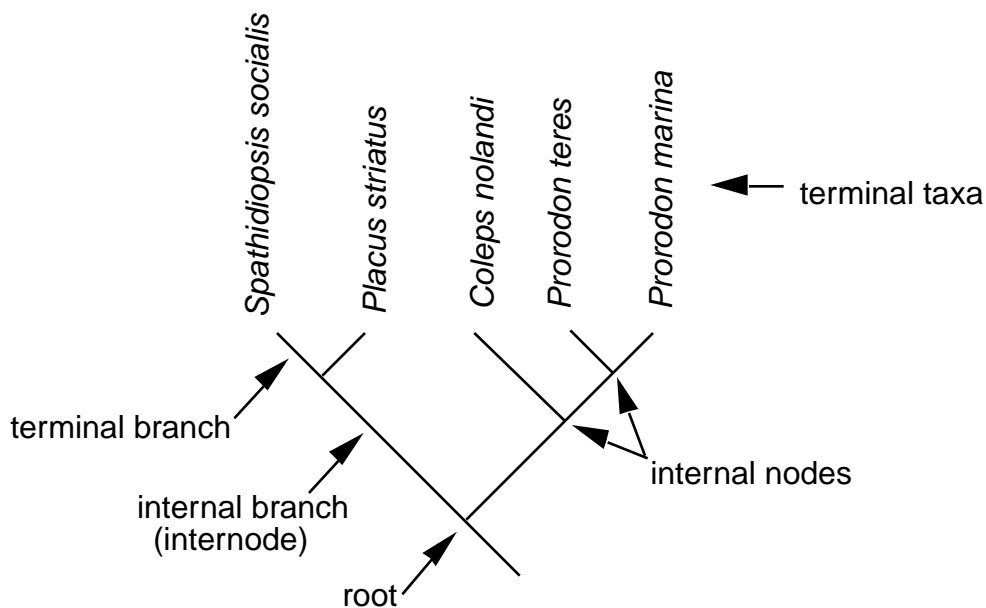
Since there are no changes between the point where **DE** branch off and **C** branches off the tree, the tree should be drawn as unresolved:



Cladograms

DESCRIPTIVE TERMS

There are several terms regarding trees that are frequently encountered and should be learned. Trees have a **root**, which is the starting point or base of the tree. The branching points are called **internal nodes**, and the segments between the nodes are called **internal branches** (or more rarely, internodes). Taxa placed at the ends of branches are called **terminal taxa** and the branches leading to them are called **terminal branches**.



WHAT A CLADOGRAM ACTUALLY SAYS ABOUT RELATIONSHIPS

The trees that result from cladistic analysis are relative statements of relationship

and do not indicate ancestors or descendants. For example in the tree above, *Prorodon teres* and *Prorodon marina* are hypothesized to be **sister taxa** and to share a more recent common ancestor with each other than with *Coleps*; but the prorodontids (*P. teres*+*P. marina*+*Coleps*) all share a more recent common ancestor with one another than with the Placidae (*Placus* + *Spathidiopsis*).

The tree does not explicitly hypothesize ancestor-descendant relationships. In other words, the tree hypothesizes that *Prorodon* and *Coleps* are related, but not that *Prorodon* evolved from *Coleps* or that *Coleps* evolved from *Prorodon*.

Descriptive Statistics for Cladograms

Various descriptive statistics have been developed to show how much homoplasy is required by a cladogram.

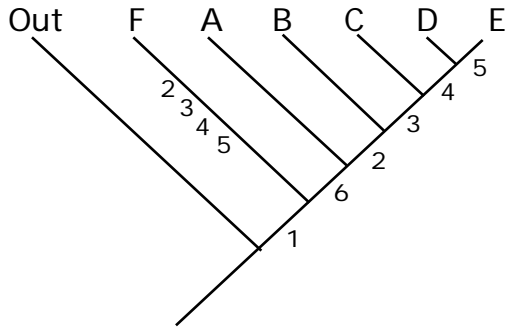
LENGTH

The **length**, or number of **steps**, is the total number of character state changes necessary to support the relationship of the taxa in a tree. The better a tree fits the data, the fewer homoplasies will be required and the fewer number of character state changes will be required. Therefore, a tree with a lower length fits the data better than a tree with a higher length. The tree with the lowest length compels us to assume fewer homoplasies and so is more parsimonious - it will be the hypothesis of taxa relationship that is selected.

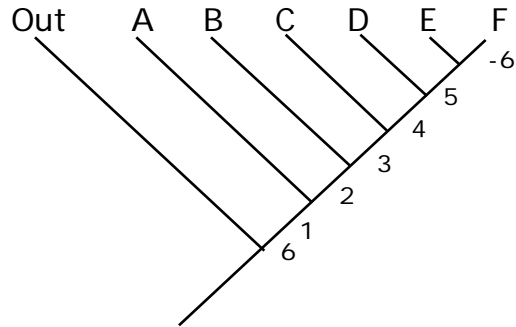
Consider the following data set of 6 taxa, 1 outgroup, and 6 characters:

	Characters					
	1	2	3	4	5	6
Outgroup	0	0	0	0	0	0
Taxa A	1	0	0	0	0	1
Taxa B	1	1	0	0	0	1
Taxa C	1	1	1	0	0	1
Taxa D	1	1	1	1	0	1
Taxa E	1	1	1	1	1	1
Taxa F	1	1	1	1	1	0

Character 6 in this data set suggests that taxa A, B, C, D, and E are in a group that excludes taxon F. The other characters suggest that taxa F and E are most closely related and that they are as a group related to taxon D; the clade with taxa F, E and D are related to C, then this group is related to B; finally, taxon A is joined to the clade containing B, C, D, E, and F:



This tree requires 10 character state changes
Length = 10



This tree requires 7 character state changes
Length = 7

The second tree has a lower length because it requires fewer homoplasious steps. In other words, it is the most parsimonious, and therefore is a better hypothesis of the relationships of the taxa.

CONSISTENCY INDEX

The relative amount of homoplasy can be measured using the **consistency index** (often abbreviated CI). It is calculated as the number of steps expected given the number of character states in the data, divided by the actual number of steps multiplied by 100. The formula for the CI is:

$$CI = \frac{\text{total character state changes expected given the data set}}{\text{actual number of steps on the tree}} \times 100$$

For example, in the data set above:

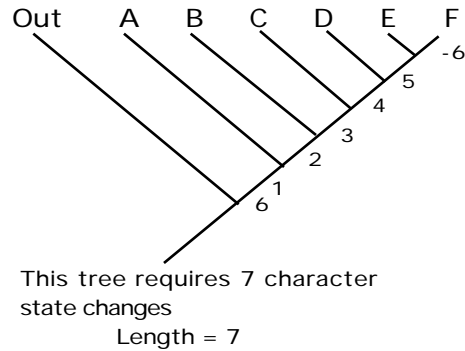
Diagnosis Data Set 1

	1	2	3	4	5	6	Character	State	Changes in Data
Outgroup	0	0	0	0	0	0		1	0 --> 1
Taxa A	1	0	0	0	0	1		2	0 --> 1
Taxa B	1	1	0	0	0	1		3	0 --> 1
Taxa C	1	1	1	0	0	1		4	0 --> 1
Taxa D	1	1	1	1	0	1		5	0 --> 1
Taxa E	1	1	1	1	1	1		6	0 --> 1
Taxa F	1	1	1	1	1	0			

Total character state changes

expected in the data set = 6

The cladogram of these data is:



The character state changes in the tree total 7 because character 6 reverses to state 0 in taxon F:

Character State Changes on Tree:

1	0 --> 1
2	0 --> 1
3	0 --> 1
4	0 --> 1
5	0 --> 1
6	0 --> 1 --> 0

The **Consistency Index** = $6/7 \times 100 = 85$

RETENTION INDEX

Another measure of the relative amount of homoplasy required by a tree is the **retention index (RI)**. The retention index measures the amount of synapomorphy expected from a data set that is retained as synapomorphy on a cladogram.

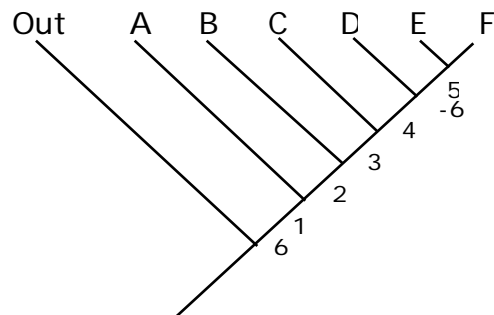
In order to describe the Retention Index and distinguish it from the CI, let's look at another example:

Diagnosis Data Set 2

	1	2	3	4	5	6	Character	State Changes in Data
Outgroup	0	0	0	0	0	0		1 0 --> 1
Taxa A	1	0	0	0	0	1		2 0 --> 1
Taxa B	1	1	0	0	0	1		3 0 --> 1
Taxa C	1	1	1	0	0	1		4 0 --> 1
Taxa D	1	1	1	1	0	1		5 0 --> 1
Taxa E	1	1	1	1	1	0		6 0 --> 1
Taxa F	1	1	1	1	1	0		

Total character state changes expected in the data set = 6

The cladogram of these data is:



This tree requires 7 character state changes
Length = 7

The character state changes in the tree total 7 because character 6 reverses to state 0 in the line leading to taxa E and F:

Character	State Changes on Tree
1	0 --> 1
2	0 --> 1
3	0 --> 1
4	0 --> 1
5	0 --> 1
6	0 --> 1 --> 0

The **Consistency Index** = $6/7 \times 100 = 85.7$

The consistency index is the same for this data set and the one used above to describe how to calculate the CI, but the data sets are not identical. In the first data set, the homoplasious character (#2) reverses in a terminal branch. In data set 2, the homoplasy defines a clade including 2 terminal branches. Unlike the first data set,

in Data Set 2 the homoplasy is informative about the branching pattern of the taxa. This additional information is measured by the retention index but is ignored by the consistency index.

To calculate the retention index:

$$\text{RI} = \frac{\text{maximum number of steps on tree} - \text{number of state changes on the tree}}{\text{maximum number of steps on tree} - \text{number of state changes in the data}} \times 100$$

The maximum number of steps on the tree is the total number of taxa with state 1 or state 0 (whichever is smaller), summed over all the characters.

The **RI** of the Data Set 1:

maximum number of steps =

	1	2	3	4	5	6	Character	Max steps	
Outgroup	0	0	0	0	0	0		1	1
Taxa A	1	0	0	0	0	1		2	2
Taxa B	1	1	0	0	0	1		3	3
Taxa C	1	1	1	0	0	1		4	3
Taxa D	1	1	1	1	0	1		5	2
Taxa E	1	1	1	1	1	1		6	2
Taxa F	1	1	1	1	1	0			

Total max steps in data set = 13

$$\text{RI} = \frac{13 - 7}{13 - 6} \times 100 = 85.7$$

But, The RI of the second data set is higher than the consistency index:

maximum number of steps =

	1	2	3	4	5	6	Character	Max steps	
Outgroup	0	0	0	0	0	0		1	1
Taxa A	1	0	0	0	0	1		2	2
Taxa B	1	1	0	0	0	1		3	3
Taxa C	1	1	1	0	0	1		4	3
Taxa D	1	1	1	1	0	1		5	2
Taxa E	1	1	1	1	1	0		6	3
Taxa F	1	1	1	1	1	0			

Total max steps in data set = 14

$$RI = \frac{14 - 7}{14 - 6} \times 100 = 87.5$$

FIT OF INDIVIDUAL CHARACTERS

It is often of interest to examine how well an individual character is fit by a tree. We can measure the length, consistency index, and retention index of individual characters.

Length

The length of an individual character is the number of character state changes or steps required to fit it to the tree. In the first diagnosis data set, character 6 has a length of 2. It changes from state 0 to 1 in the branch leading to the ingroup taxa (step 1) and changes again from state 1 to 0 in the line leading to taxon F.

Consistency Index

The consistency index for an individual character is the number of advanced states it has in the data set (i.e., the number of state changes we expect) divided by the number of steps in the tree multiplied by 100. In the first diagnosis data set, character 6 has only one advanced state (state 1), but it actually requires 2 steps to fit the cladogram. The CI of this character is:

$$CI = \frac{1}{2} \times 100 = 50.0$$

Retention Index

The retention index for an individual character is the number of taxa with states 1 or 0 (whichever is higher) (this is the maximum steps for the character) take away the number of steps the character makes in the tree divided by the maximum steps for the character take away the number of state changes we expect multiplied by 100. In the first diagnosis data set, character 6 has two taxa with state 0 - therefore 2 is

the maximum steps. The RI of this character is:

$$RI = \frac{2 - 2}{2 - 1} \times 100 = 0.0$$

Optimality Criteria

When trees are constructed, parsimony chooses the hypothesis that requires the fewest number of steps. This means that the number of steps between each state of a character must be determined before the cladogram is built. The criterion used to set the number of steps between each state is called the **optimality criterion**. There are several different optimality criteria, the pros and cons of which are described below.

WAGNER OPTIMALITY

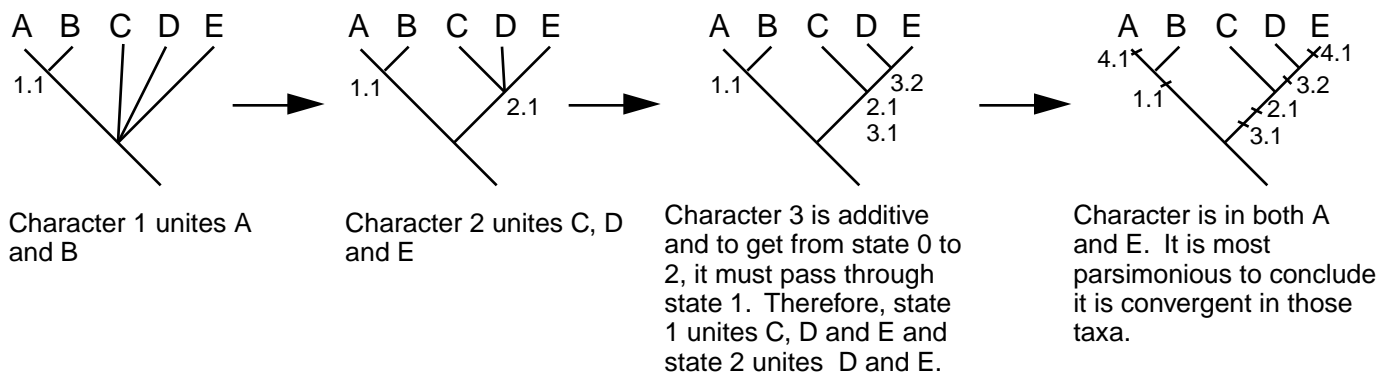
Wagner optimization, formalized by Farris (1970) and based upon the work of Wagner (1961), is one of the simplest optimality criteria.

1. Characters are allowed to reverse so that change from 0→1 costs the same number of steps as 1→0.
2. Characters are additive so that if 0→1 is 1 step, and 1→2 is 1 step, then 0→2 must be 2 steps.

To see how this works, consider the following data set:

```

Out 0000
A   1001
B   1000
C   0110
D   0120
E   0121
  
```



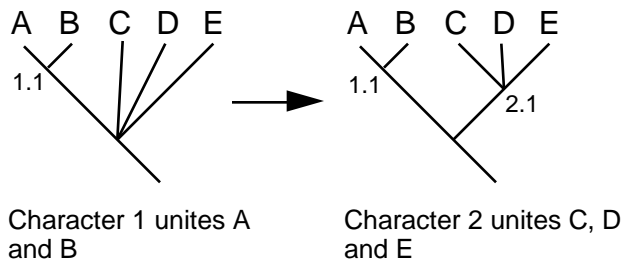
FITCH OPTIMALITY

Fitch optimization (Fitch 1971) is similar to Wagner method in that characters are reversible, but differs by allowing characters to be non-additive.

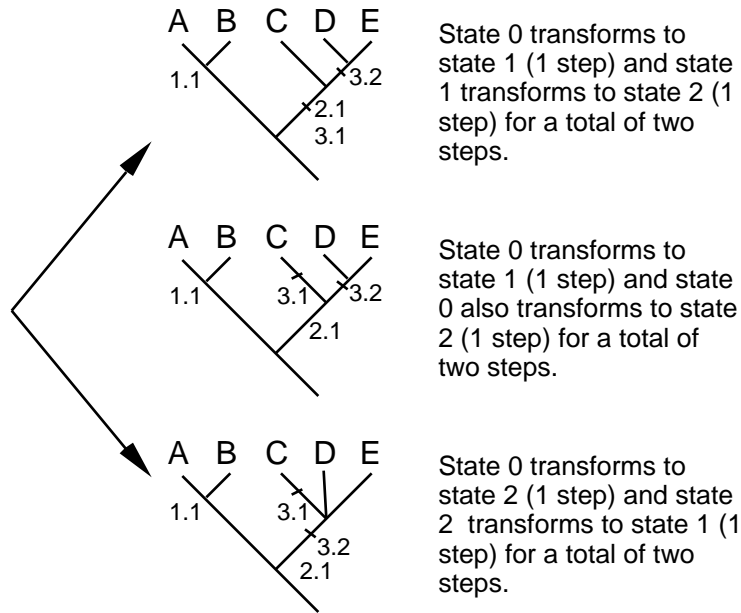
1. Characters are allowed to reverse so that change from $0 \rightarrow 1$ costs the same number of steps as $1 \rightarrow 0$.
2. Characters are nonadditive so that if $0 \rightarrow 1$ is 1 step, and $1 \rightarrow 2$ is 1 step, then $0 \rightarrow 2$ is also 1 steps.

To see how this works, consider the data set above.

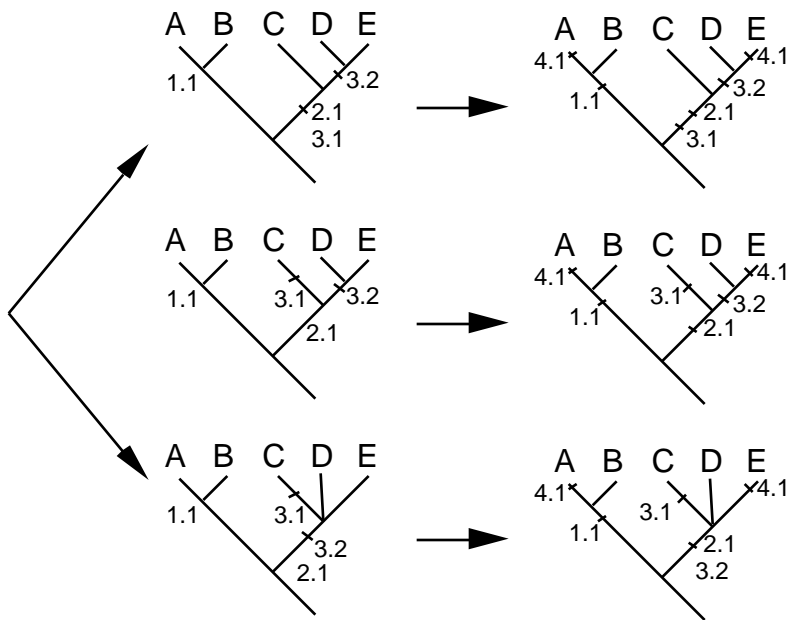
Analysis of two-state characters is the same in Wagner and Fitch optimality, therefore the analysis of the first two characters is the same as above:



Character 3 demonstrates the difference between the Fitch and Wagner optimality. In the Fitch method the states are not additive so that the following transformations all take two steps:



Character 4 is in A and E, and it is most parsimonious to assume it arose as a convergence in these two taxa:

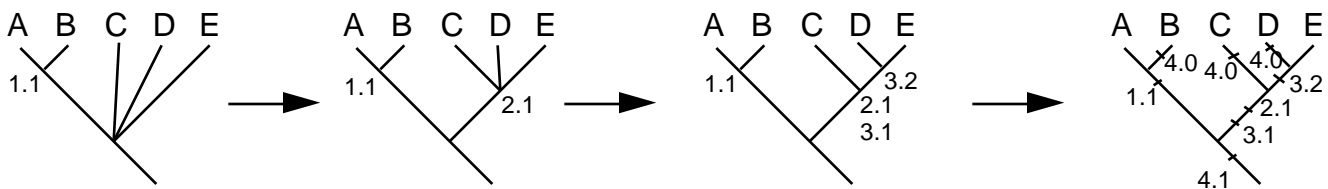


Fitch optimality always results in more equally parsimonious trees than Wagner optimality.

DOLLO OPTIMALITY

L. Dollo (1893) noted that evolution rarely reverts to an earlier specialized form. He called this the “Law of Phylogenetic Irreversibility” but it is usually known as Dollo’s Rule. An example would be the evolution of the amniote forelimb. The exact arrangement of humerus, radius, ulna, wrist bones, and digits would have only evolved once. It would be impossible for such a complex structure to have evolved a second time so all homoplasy must be considered to be secondary loss.

Using our sample data set, the difference between this method and Wagner optimality would be in the treatment of character 4. This homoplastic character, is made to fit the tree as secondary loss in B, C and D:



The problem with this method is that it requires us to assume a model of evolution.

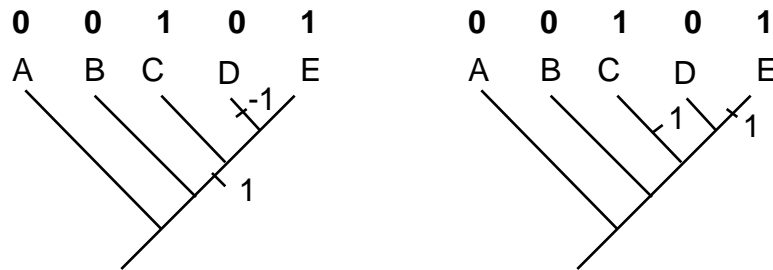
CAMIN-SOKAL OPTIMALITY

Camin-Sokal optimization (Camin and Sokal, 1965) holds that once a state has been acquired it may never be lost. Thus any homoplasy must be accounted for by multiple origin. With the sample dataset, the result would be the same as with Wagner optimality. The problem with this method is that it requires us to assume a model of evolution.

OPTIMIZING CHARACTERS ON A CLADOGRAM

As we have seen, character state changes are determined when the tree is built and where the character changes occur depends upon the optimality criterion used. Sometimes a character may be optimized in more than one equally parsimoniously

way. If the following character is optimized using either Wagner or Fitch optimality criterion, two optimizations are possible:



The first tree is has an **ACCTRAN** (ACCelerates the TRANSformation of characters on a tree) optimization- the character is optimized as close to the root as possible.

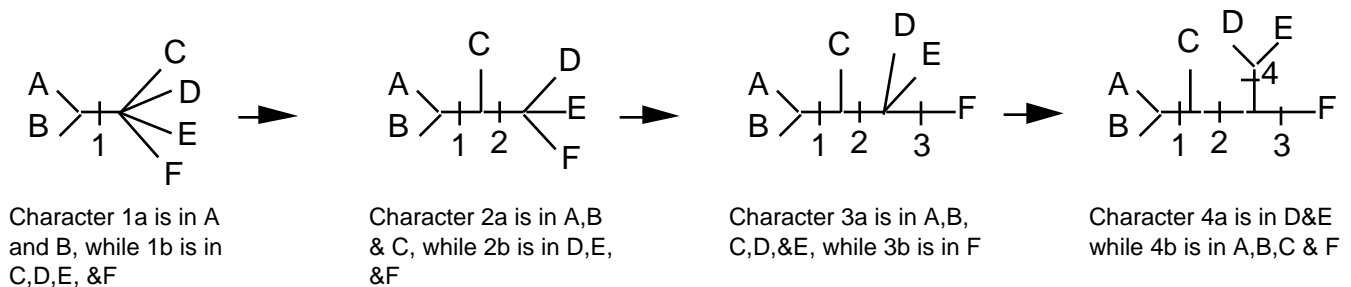
The first tree is has a **DELTRAN** (DELays the TRANSformation of characters on a tree) optimization - the character is optimized as far from the root as possible.

Rooted and Unrooted Trees

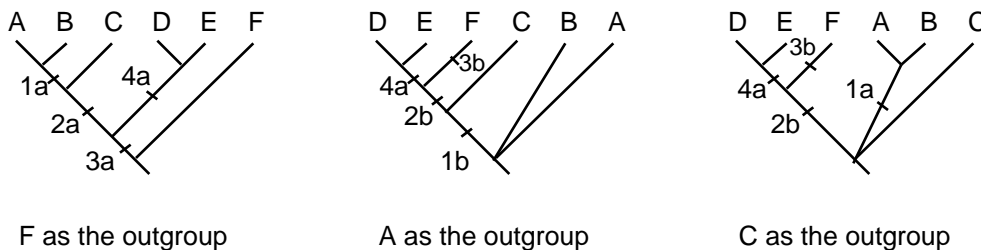
When the Wagner and Fitch optimality criteria are used, the purpose of the outgroup is simply to create the root of the tree. This can be seen if we build an unrooted tree or **network**. Consider the following dataset:

A aaab
 B aaab
 C baab
 D bbaa
 E bbaa
 F bbbb

Rather than creating a cladogram, create a network based on shared states:



The final network requires four steps. No matter which taxon is treated as the outgroup, the resulting tree will have just four steps:



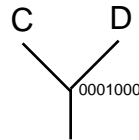
Not all possible trees are shown, but all have just four steps.

Because a dataset will always result in fewer equally parsimonious networks than trees, many computer programs search for equally parsimonious networks to save computation time.

Searching for Most Parsimonious Trees

Building cladograms becomes more complicated as datasets become larger or have more character conflicts. Consider the following data set and the resulting cladogram (calculated here by the Wagner Method):

	1	2	3	4	5	6	7
A	1	0	0	1	1	1	1
B	0	1	0	0	1	0	1
C	0	0	1	1	0	1	0
D	0	0	0	1	0	0	1



Attach B:

```

B 0100101
C 0011010
  111111      6 steps
  
```

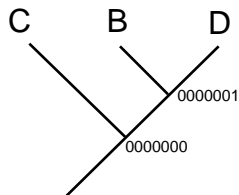
```

B 0100101
D 0001001
  1 11      3 steps
  
```

```

B 0100101
CD 0001000
  1 11 1    4 steps
  
```

Therefore, B should be attached to D:



Now attach A

```

A 1001111
C 0011010
  1 1 1 1      4 steps
  
```

```

A 1001111
D 0001001
  1 11      3 steps
  
```

```

A 1001111
B 0100101
  11 1 1           4 steps

```

```

A 1001111
BC 0000001
  1   111       4 steps

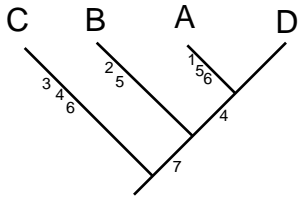
```

```

A 1001111
BCD 0000000
  1   1111      5 steps

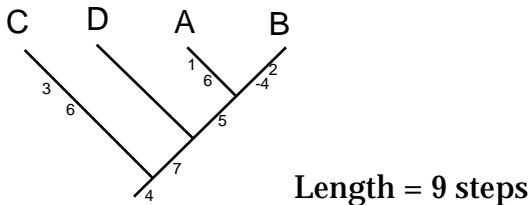
```

Therefore A should be attached as sister-taxon to D.



The cladogram is complete and has a length of 10 steps.

Everything seems OK, but move the branch leading to B so that A and B are sister taxa and determine the length of the new tree:



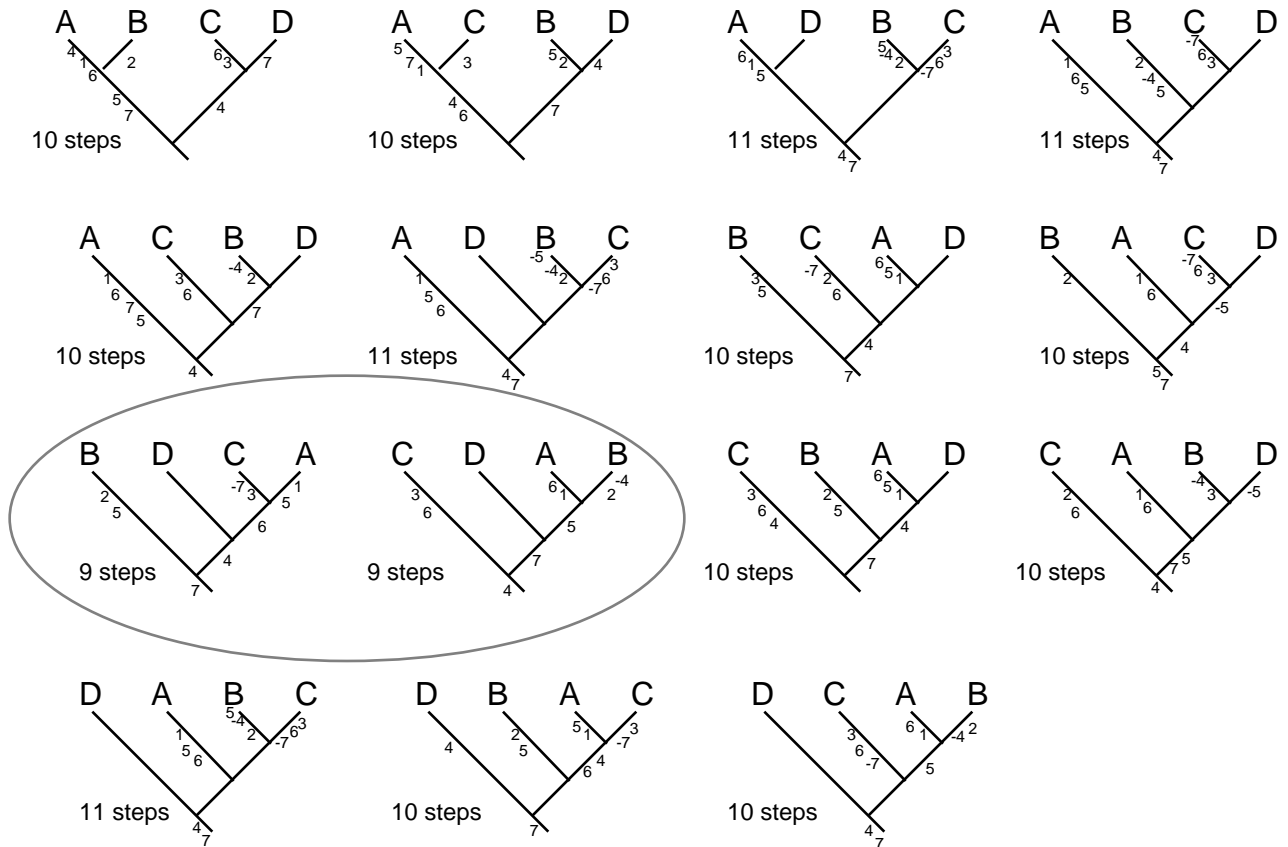
This a better (more parsimonious tree). What has happened here? We did the method correctly, but we didn't end up with the correct cladogram for the data.

The Wagner method is a procedure that works by step-wise addition of taxa such that each taxon is added where it optimally fits gives the tree at that time. As more taxa are added, a better placement for the taxon might appear but, because it is already attached to the tree, that possibility is not apparent.

EXHAUSTIVE SEARCH

A solution to finding the most parsimonious tree is simply to check all possible

trees. The easiest way to do this is to start by checking all the fully resolved trees, then if the shortest tree has branches not supported by any characters, collapse those branches to get the correct unresolved tree. For the data set above, the possible trees are:



By checking every possible tree, we find that there are two most parsimonious trees (circled above) that are one step shorter than the tree found by the step-wise Wagner procedure.

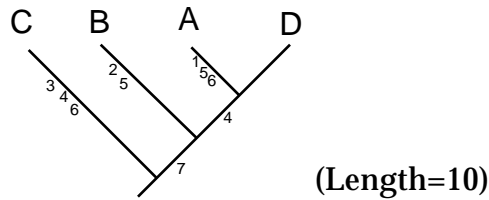
Unfortunately, this method is only feasible for data sets with fewer than 11 taxa, because the number of possible trees quickly becomes enormous (for just 7 taxa there are 945 trees, 2×10^6 trees for 10 taxa, and 2×10^{20} for 20 taxa).

BRANCH AND BOUND SEARCH

The branch-and-bound method saves effort by only checking trees that are likely to be the shortest. This is how it works:

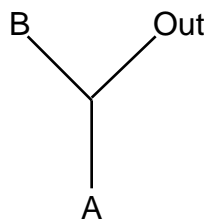
1. First a tree is constructed using a method such as the Wagner Method:

	1	2	3	4	5	6	7
A	1	0	0	1	1	1	1
B	0	1	0	0	1	0	1
C	0	0	1	1	0	1	0
D	0	0	0	1	0	0	1
Out	0	0	0	0	0	0	0

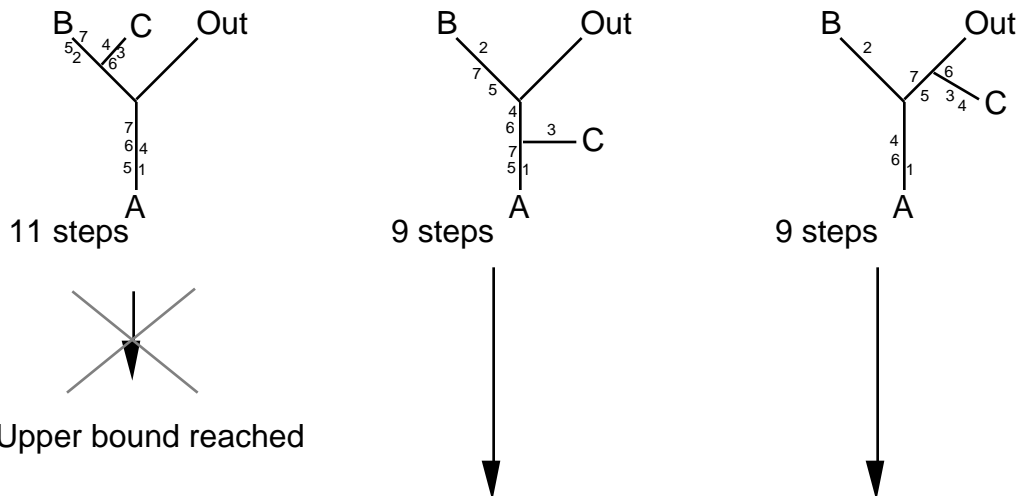


This tree length is set as the upper bound -- we know the most parsimonious tree(s) must be of length 10 or lower.

2. Two taxa are chosen (it does not matter which two) and connected to the outgroup into the only unrooted tree possible:



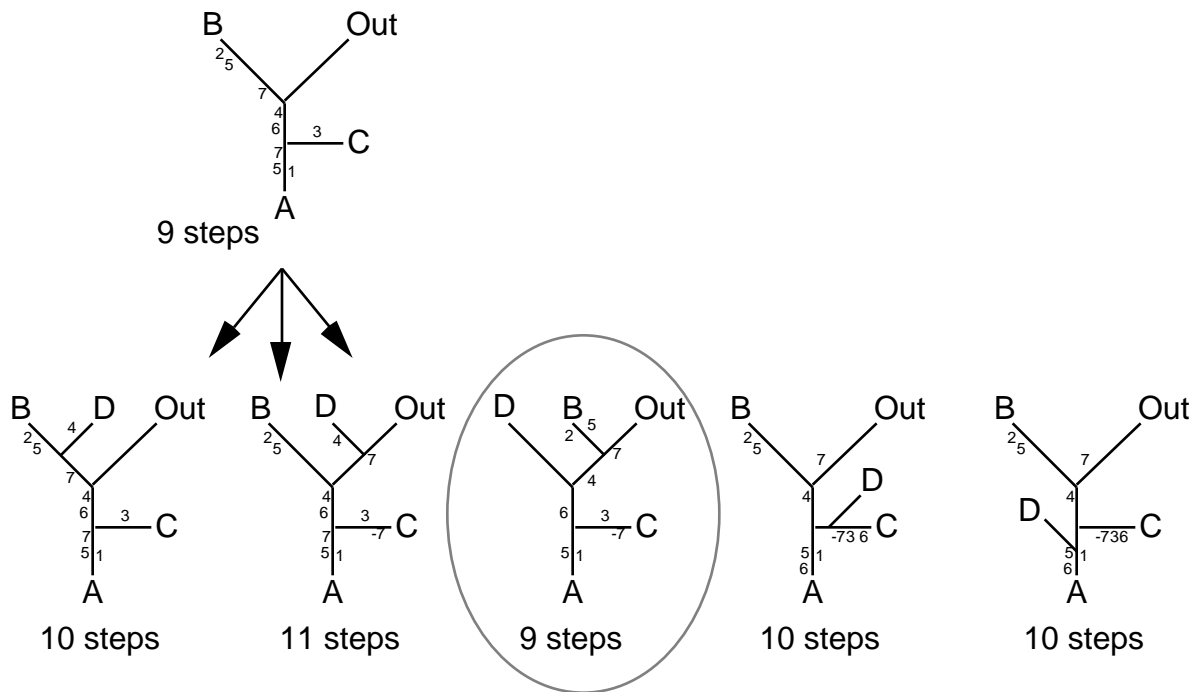
3. A third taxon is added to every possible position on the network, and the number of steps each requires is noted:

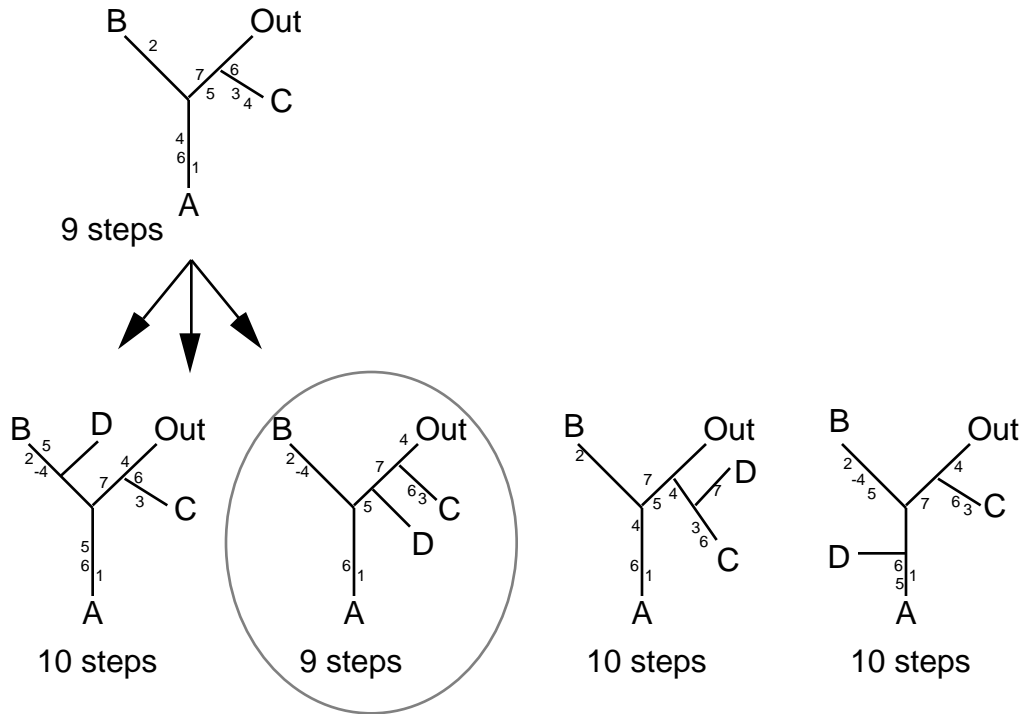


One of the unrooted trees has already reached 11 steps, which is greater than the

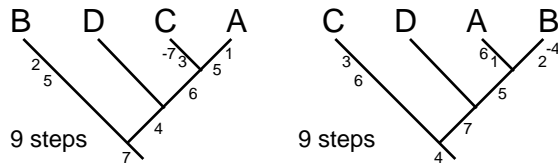
upper bound. We don't have to continue to check trees based on this pathway because when the next taxon is added, the length of those trees will not be as short as the most parsimonious tree, and so will be rejected. Because there is no reason to examine these pathways further, we have reduced the number of trees to be checked.

4. The fourth and final taxon is added to every possible position on the network, and the number of steps each requires is noted:





Using the outgroup to form the root, we get the two most parsimonious trees:



Although branch-and-bound methods require checking fewer trees than an exhaustive search, it is still time consuming and may be impractical for large, messy datasets. For these datasets, inexact heuristic methods are used.

BRANCH SWAPPING

In branch swapping, the branches of a tree are rearranged to search for a shorter topology. In the example above, branch swapping was performed and a shorter tree found when the branch leading to B was moved so that A and B became sister taxa. In branch swapping not every possible rearrangement is checked -- that would be the same as checking every possible tree (an exhaustive search). Instead, computer programs make shortcuts:

1. Local branch swapping (also called nearest neighbor interchange) moves swaps the branches adjacent to each other on an internal branch.

2. Global branch swapping clips a cladogram into two or more subcladograms, and then rearranges the subcladograms into new trees. This can be done by:
 - a. Subtree pruning and regrafting - a sub cladogram is clipped off the main cladogram, then reattached to each branch in turn.

 - b. Tree bisection and reconnection - the clipped subclade is re-rooted before it is reconnected to each branch. All possible bisections, rerootings and reconnections are evaluated.

The more extensive the branch swapping, the greater the chance of uncovering the shortest tree, but the longer it takes to do the calculations. Balancing the need for precision in finding the shortest tree against a reasonable amount of computation time is one of the most difficult computational problems for systematists.

Classifications

One of the major products of systematics is the formal classification system for species. These names are handles by which information and communication about organisms and diversity are conveyed.

The system of classification that we use today was originally described by **Carolus Linnaeus**.

The groups into which organisms are placed are referred to as taxa (singular, taxon). The taxa are arranged in a hierarchy. The broadest taxa contain a large number of organisms that share very fundamental characteristics. Each broad taxon includes many smaller, more inclusive taxa (each of which contains organisms that share increasingly more specific characteristics). The levels in the hierarchy are:

Kingdom
Phylum (plural, phyla)
Class
Order
Family
Genus (plural, genera)
Species

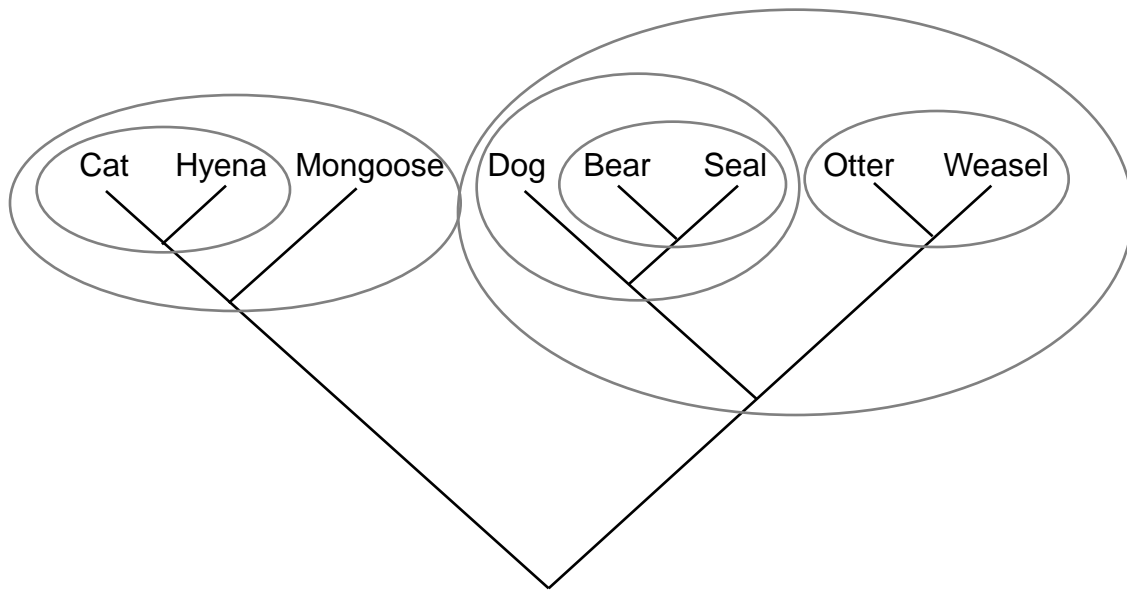
A biological classification should

- (1) summarize the characteristics of organisms efficiently (in other words, have high information content), and
- (2) reflect real groups (in other words, reflect evolutionary relatedness or phylogeny).

Therefore, we can use our phylogenetic trees to not only tell us about relatedness, but also to help us create a classification.

MONOPHYLY

One of the tasks of a systematist is to convert the tree into the formal hierarchical Linnean classification by giving groups that share a common ancestor a formal taxonomic name. Such groups are called monophyletic taxa and they are recognized because they share unique derived characters. The tree shows several sets of most closely related taxa, that are nested within larger sets. Because Linnean categories are also interested, these sets of taxa can be converted into Linnean categories. The monophyletic groups are:



The formal taxa for these groups are given below. Notice, however, that not all of the groups have been named. For example, there is no formal taxon for the group that includes dogs, bears, and seals:

Order Carnivora

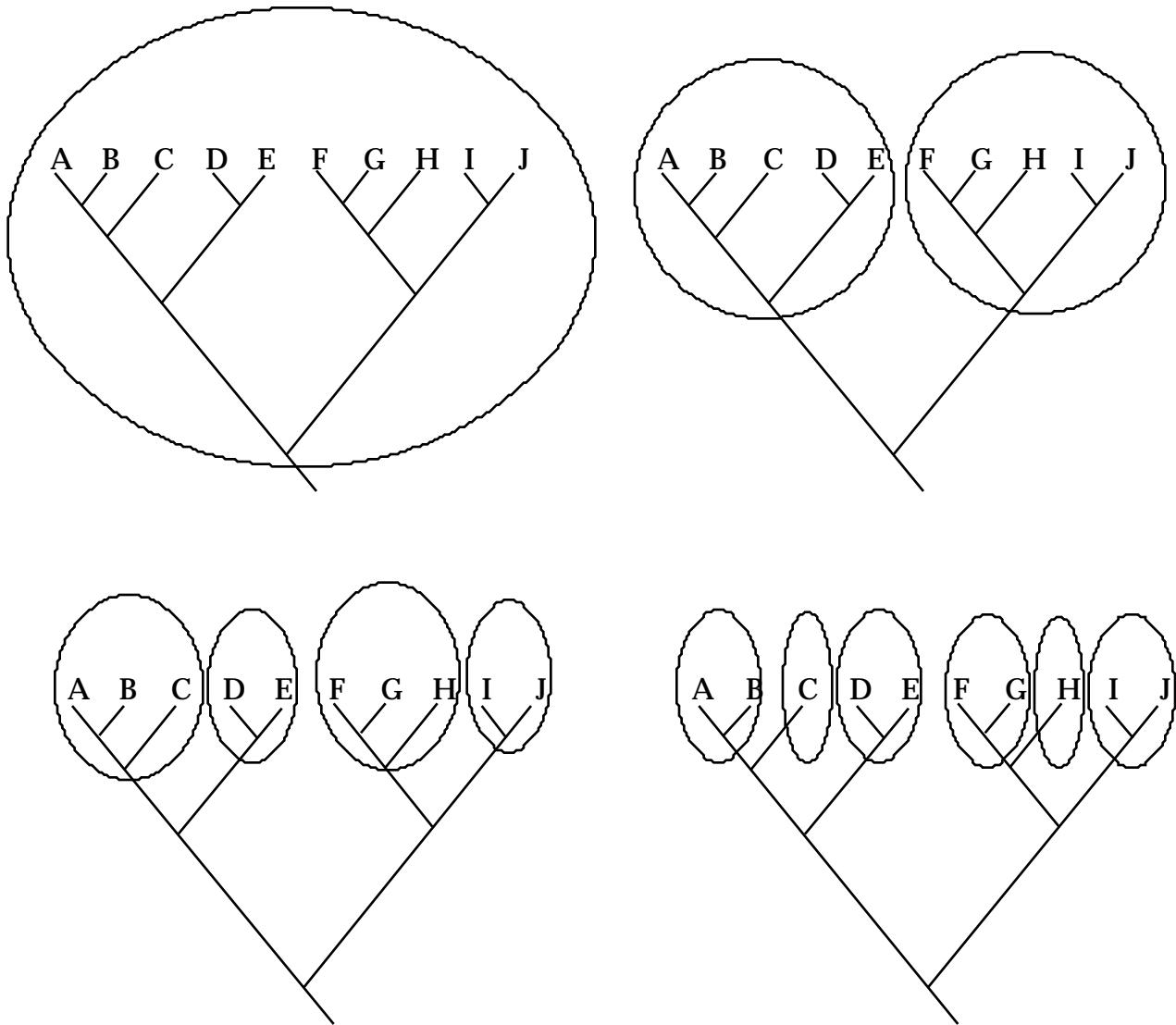
Suborder Feliformia
Superfamily Feloidea
Family Felidae
Genus *Felis* - cat
Family Hyaenidae
Genus *Hyaena* - striped hyena
Family Herpestidae
Genus *Herpestes* - mongoose

Suborder Caniformia
Superfamily Arctoidea
Family Canidae
Genus *Canis* - dog
Family Ursidae
Genus *Ursus* - brown bear
Family Phocoidea
Genus *Phoca* - harbor seal
Family Mustelidae
Genus *Mustela* - weasel
Genus *Lutra* - otter

ARBITRARINESS IN THE FORMAL CLASSIFICATION SCHEME

Lumpers and Splitters

The flexibility of the classification system, however, also makes it somewhat arbitrary in that there are no clear guideline as to what is the cut-off for recognizing one large taxon versus two smaller ones.



Some biologists (called “**lumpers**”) focus on and stress similarities held in common by the organisms being studied and so tend to group several species into a single genus. Other biologists (called “**splitters**”) focus on and stress differences

between the species and so tend to divide the species into several different genera.

WHY DO CLASSIFICATION SCHEMES CHANGE?

New Data -

New technologies constantly give rise to new sources of character information. New information reveals new similarities and differences among taxa that cause us to revise the placement of a taxon in a tree or to choose to lump or split a taxon within an existing classification.

New Taxa -

As previously unknown species are discovered, classifications will also need to be revised to reflect their placement. This will undoubtedly have a large impact on existing classification schemes because, at this time, we cannot say how many more species exist on earth waiting to be discovered.

Misinterpreted data

Finally, new studies occasionally lead to the discovery that features used to group species into a taxon are actually convergent or nonunique characters. When this happens, the old taxon is abandoned and a new monophyletic taxon is created in its place. There are two instances where this occurs:

Polyphyly - Occasionally, new studies lead to the discovery that features used to group species into a taxon are actually convergent characters. The taxon is then known to be polyphyletic (taxa that do not share a recent common ancestor and were grouped on the basis of homoplasy).

Paraphyly - Inevitably some plesiomorphic characters are incorrectly interpreted to be synapomorphies and a paraphyletic taxon is created.

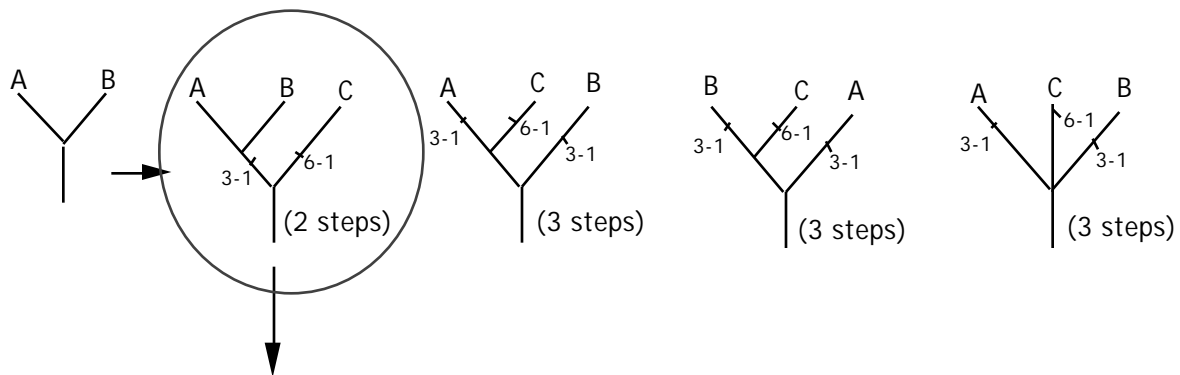
Multiple Equally Parsimonious Trees

Although some data sets will yield a single most parsimonious tree, other data sets give two, three, or even hundreds of equally parsimonious trees.

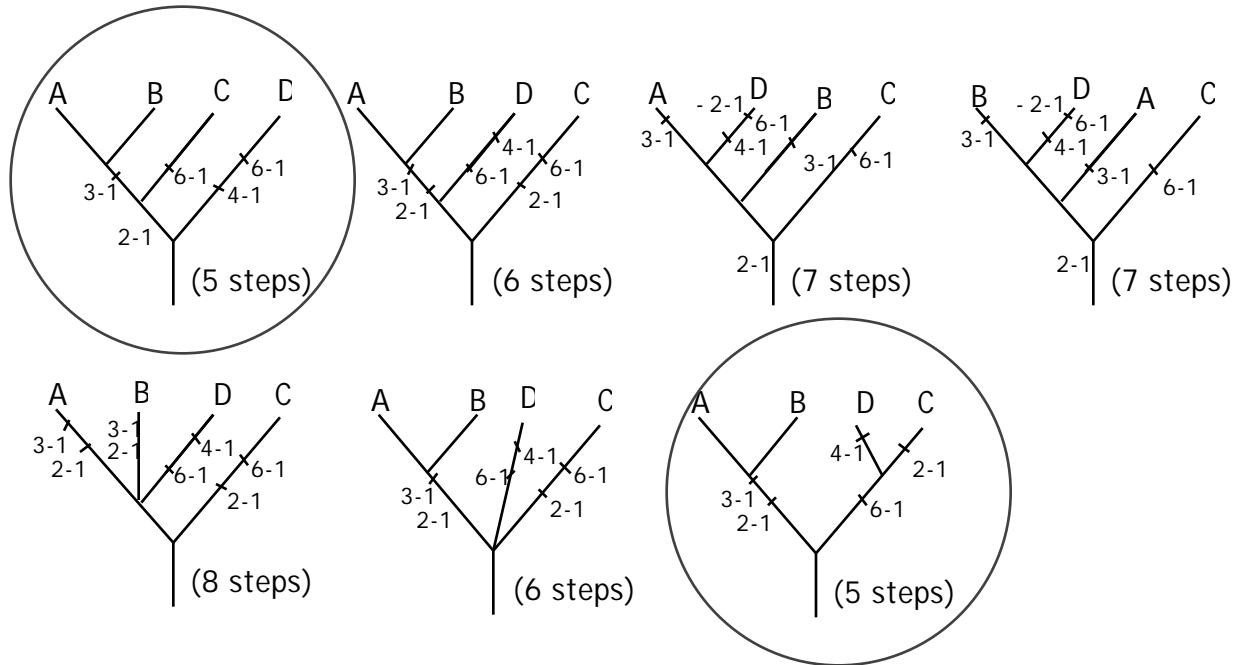
Consider the following example:

Outgroup	000000
A	111000
B	111000
C	110010
D	100101
E	100110
F	100111

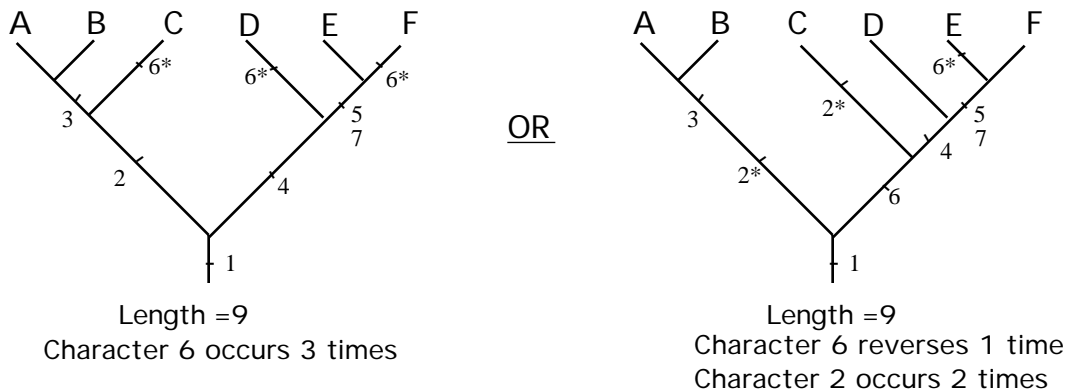
Using the Wagner method, we add one taxa at a time to the place where it requires the fewest steps. In the example below only the relevant characters that vary between the taxa are considered:



Use the tree with 2 steps, and determine the placement of taxon D:



There are two equally parsimonious placements of taxon D - as the sister group to taxon C and as the sister group to the ABC clade. In the next step, the placement of taxon E must be calculated for both trees. In the end two equally parsimonious cladograms are found:



There are essentially two different strategies for coping with multiple solutions. The worker may accept that some parts of the classification cannot be resolved satisfactorily at present and may therefore concentrate on only those parts that are found consistently in all classifications. On the other hand, the worker may justify choosing one tree over another. The former approach uses consensus

methods, the latter generally involve some sort of weighting of the evidence.

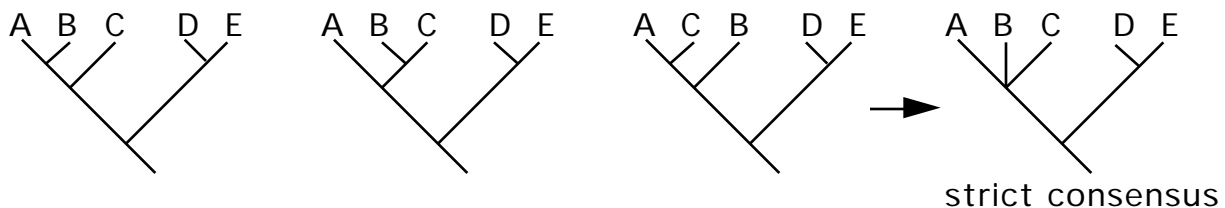
CONSENSUS TREES

Consensus trees are used to show the information about taxa relationships that all the equally parsimonious cladograms have in common. There are several ways to form consensus cladograms: Adams, strict, combinable components, and majority rules consensus trees.

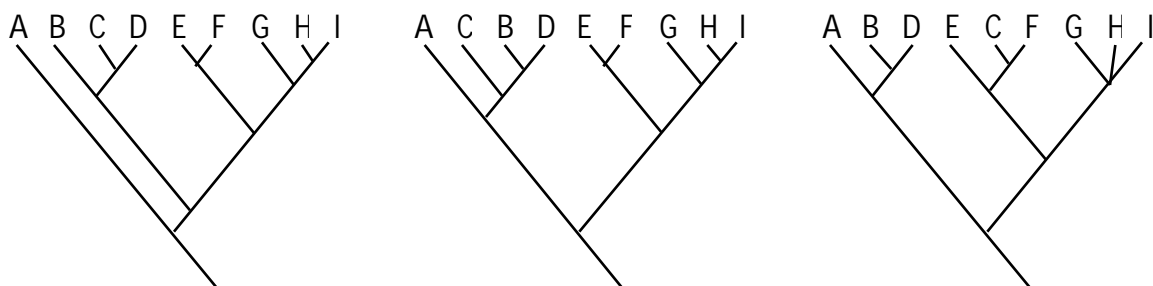
Strict Consensus

The most conservative method is strict consensus and is derived by combining only those components from a set of trees that appear in all of the original trees.

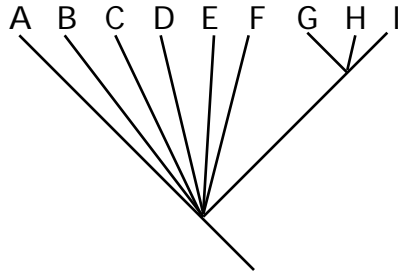
The following three trees agree that A+B+C form a clade but differ in which pair of taxa are more closely related. The strict consensus reflects only that information that all three trees unambiguously share:



The highly conservative nature of strict consensus sometimes means that one may be left with little resolution. Consider the following three trees:



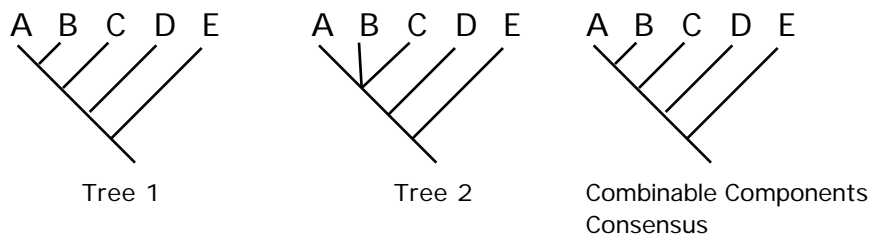
The strict consensus for these trees is:



The rationale for strict consensus is that it includes only those components that are totally unambiguous and about which the data are absolutely clear. Other consensus methods allow groups that are not fully supported or are supported ambiguously. For this reason, Nixon and Carpenter (1996) argued that all other methods result in “compromise trees” rather than “consensus trees.”

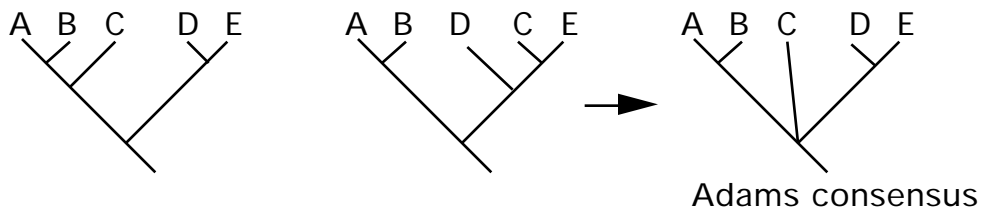
Combinable Component

A combinable components consensus tree is similar to a strict consensus tree, except that it will combine (or include) those clades that are **not contradicted** by all of the trees (Bremer, 1990). Nonconflicting components occur when at least one of the original cladograms has an unresolved component (or polytomy). It is often better resolved than a strict consensus tree and it may result in a tree with greater resolution than one of the original trees if one of the original trees is poorly resolved:



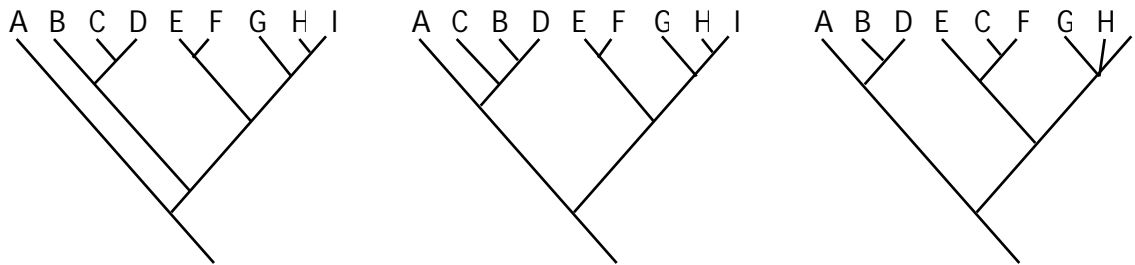
Adams Consensus Cladograms

The Adams consensus tree places taxa that conflict at the node all the conflicting positions have in common (Adams, 1972). This method is very useful when one taxon is causing all of the tree conflict:

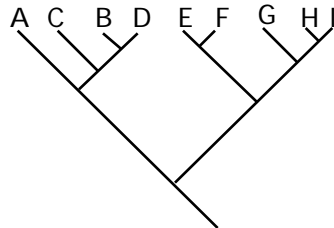


Majority Rules Consensus

The most frequent placement of a taxon in all the trees is its placement in the consensus tree (Swofford, 1991):

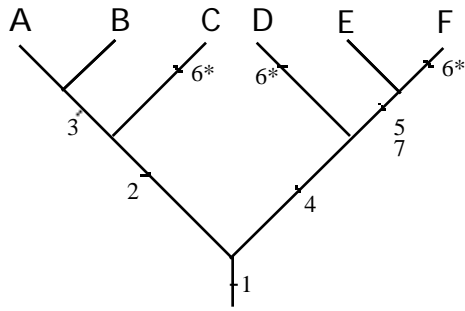


The majority rules consensus tree is:



Problems With Consensus Trees

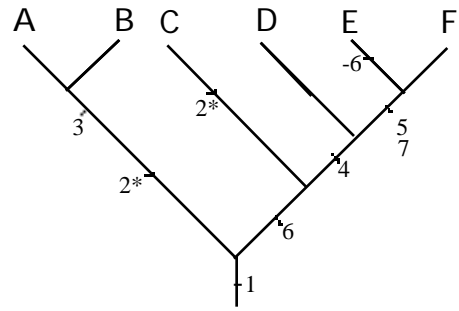
1. A consensus tree will always require more steps (have a longer length and lower consistency index) and will, therefore, be less parsimonious than any of the trees from which it was formed (Miyamoto, 1985, Carpenter, 1988):



Length = 9

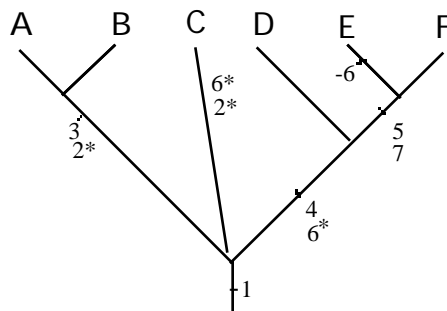
Character 6 occurs 3 times

OR



Length = 9

Character 6 reverses 1 time
Character 2 occurs 2 times



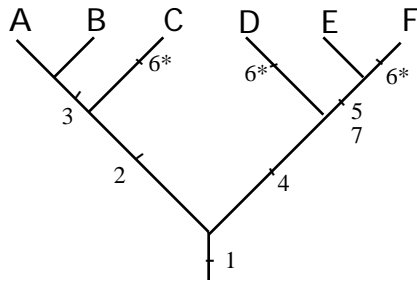
Length = 10

Strict Consensus Tree

2. Consensus trees result in unresolved branching points. Unresolved nodes imply many relationships of the taxa, but not all of these relationships may be possible given the original data set and cladograms.

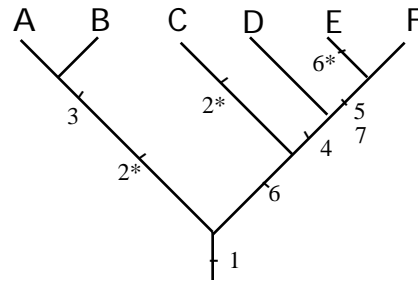
SUCCESSIVE WEIGHTING

When many equally parsimonious trees are produced, other criteria than strict parsimony may be used to choose among them. One such criterion is to choose the cladogram which requires the fewest number of characters to have homoplasies. For example, from the data set given at the beginning of the handout, two cladograms were constructed:



Length = 9
Character 6 occurs 3 times

OR



Length = 9
Character 6 reverses 1 time
Character 2 occurs 2 times

Although both trees have the same length, the first tree requires only one character to be homoplasious. Therefore, the first tree may be chosen over the second because it has both the shortest length and the fewest changing characters whereas the second tree has only the shortest length.

The tree with the fewest characters having homoplasies can be determined using a procedure called **successive weighting**. In this procedure, the *best fits* of the characters are used to calculate weights. If the best fit of a character is to have a consistency index of 100, then the character will get a high weight. If the best fit of a character is a consistency index of 33, the character will get a low weight. The weighted characters are then used to create a new cladogram. If many equal cladograms are found with the weighted data, recalculate new weights and a new tree. Continue the process until either a minimum number of cladograms are obtained.

Recently Goloboff (1993) has proposed a method that is not iterative but instead calculates weights simultaneously as the tree is built. In this method, as a character is added in tree building, its consistency index and retention index is calculate for every possible branch to which it can be added. The weight is calculated from the best of these, and then the weighted character is used to construct the tree.

Problems With Successive Weighting

1. Different sets of initial weights may lead to different final solutions. This is a problem because there is, as yet, no way to choose between different starting points (using equal weights as a starting point is arbitrary -- things are still weighted, just equally).

Question:

Turner and Zandee (1995) attacked successive weighting on the following grounds: Homoplasy is no reason to decrease our confidence in a character, and therefore weighting based on homoplasy is not justified. What do you think?

Characters

All comparison -- biological or not -- involves the correspondence of similarities. The evidence that organisms are related comes from similarities between them. The observable parts, or attributes, of an organisms which can be examined for similarity or difference are called **characters**. The alternative forms of a character are called **character states**.

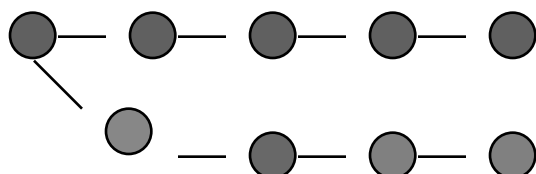
For example, suppose an entomologist looks at the attributes of the abdomens of bee species for similarities and notes that some have red abdomens while others have black abdomens. In a systematic study, the entomologist may choose *abdomen color* as a character with *red* or *black* as the character states.

Characters have alternative states because, in evolution, characters have two options: they can remain the same and be passed on genetically from ancestor to descendant unaltered:



OR

they can change in one species and be transmitted in a new form to the descendants:



Characters and their states form interested sets.

Coding Characters

There is a conventional style used to describe characters and the states. This style has the added advantage of being easy to convert to a computer file. It consists of a list of the characters used and a table showing the distribution of the character

states in the organisms.

1. **Character List** - Most Systematic papers list the characters examined and describe the character states found in the organisms in a table.

2. **Distribution of states** - which character states are found in which organisms are usually given in a matrix with the number corresponding to the character is listed across the top and the organisms are given down the side. The state for an organism is given where the name of the organism and the character intersect.

HOMOLOGY

The concept of homology is fundamental to any attempt to reconstruct phylogeny using characters. The foundations of the modern concept of homology were laid down by Owen in 1843¹ when he first made a clear distinction between what was meant by analogous and homologous anatomical parts:

A homologue is the same part or organ in different animals under every variety of form and function. or Structures in different organisms that require or deserve the same name.

An analogue, on the other hand, were different organs with the same function.

To Owen it meant the variant of the same part of the idealistic archetype from which a group of organisms was created to deviate, each in its own way.

Information was continuous, in God's mind from the archetype to its various imperfect manifestations.

¹At least in English. He may have been influenced by Goethe (1790):
“When now the plant vegetates, blooms, or fructifies, so it is still the same organs which, with different destinies and under protean shapes, fulfill the part described by nature.”

Owen later distinguished 3 types of homology:

1. **Special homology** - definition above
2. **General homology** - similarity to an archetype
3. **Serial homology** - repetition of corresponding structures in an organism.

Owen called these repetitive parts homotypes and, therefore, serial homology is sometimes referred to as homonymy.

With the general acceptance of evolution, the concept of homology took on a new meaning, usually called Darwinian homology:

Two structures in different organisms were considered to be the same structure (and thus deserve the same name) if they could be traced to the same structure in the common ancestor of the two organisms.

Darwin did not himself advance this formally as a definition - he only produced an explanation for why there were such similarities, but it was soon so taken by evolutionary systematists and has been adopted by a majority of them (Simpson, 1961). This definition is found in Mayr (1969) and Simpson (1961) and similar books. Thus, analogy is functional similarity only; homology is resemblance due to inheritance from common ancestor.

But:

Today the pendulum has swung so far from the original implication in homology that some recommend that we define homology as any similarity due to common ancestry, as though we could know the ancestry independently of the analysis of similarities.

Boyden (1947)

Most systematists today would argue that recognizing homology is a 2-step process involving both similarity and congruence. In the first step, similarity is used to postulate homology. The similarity criteria are sometime called Remane's

criteria because he first listed criteria that did not make assumptions about evolution and discussed them in detail (Remane, 1956). The list was extensively discussed and expanded by Simpson (1961):

- 1. general similarity
- 2. similar position in the body and adjacent to similar structures
- 3. similar ontogeny (similarity of the structure in both form in the embryo and how it forms)
- 4. similar genetic basis
- 5. similar even if the animals use them differently (they have different functions)
- 6. similar even in minute detail
- 7. linking intermediates between the two dissimilar forms
- 8. similar in complex structure (intricately complex systems are unlikely to have evolved more than once).

Similarity is not sufficient to recognize homology alone because (1) homoplasious characters may also pass these criteria; and (2) some features may have diverged so much that they are no longer similar, although they are homologous (e.g., inner ear bones in mammals and jaw bones in fish).

So a second test² or step - congruence, or agreement with other characters is used (Patterson, 1982). Congruence is one of the most decisive tests we use in biology - we expect other information to confirm or refute our initial hypothesis.

Homologous similarities are inferred inherited similarities that define sets of organisms as members of a group. Viewed in this way, we can see that homology is simply a character that defines a branch on the tree. In other words it is synonymous with synapomorphy (Hennig, 1966; Nelson, 1970; Wiley, 1975; Cracraft, 1978; Gaffney, 1979; Patterson, 1982).

Can we be misled using the similarity + congruence test and mistake a

²The notion of testing homologies is a fairly recent one; stemming from the introduction of Popperian philosophy into systematics.

convergent feature for a homology? Undoubtedly so. The establishment of homology is never absolutely 100% certain, but to insist on that fact as a deficiency of systematics would tend to negate the value of any and all scientific endeavor.

MISSING AND INAPPLICABLE CHARACTERS

Not all characters are available for all of the organisms in the study. This is generally because the characters are either missing (specimens or the relevant parts of specimens are missing) or inappropriate (the character is not relevant for some of the taxa). **Missing or inappropriate** characters can be represented by a question mark (?) or a dash (-). The question mark or dash signifies that the character should not be used as evidence for classifying that organism. While this sounds simple enough, analysis of datasets containing question marks is far from straightforward. Furthermore, missing characters should be treated differently from inappropriate characters.

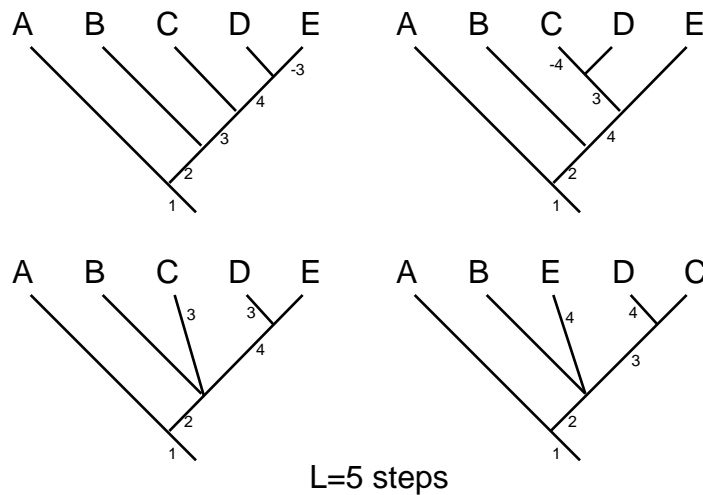
MISSING CHARACTERS

Characters might be missing for a number of reasons including inadequate sampling of all life history stages or genders for some taxa, or poorly preserved or fragmentary specimens (e.g., as in fossils). The “?” entries for such cases might represent any one of the existing states observed in other taxa. For example, for a binary character, the question mark would in reality be either a “0” or a “1”. When constructing a tree, we should allow for both possibilities when searching for the most parsimonious cladogram.

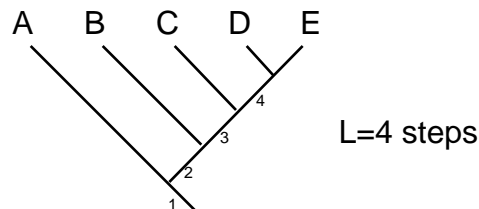
Consider the following data set:

Out	0000
A	1000
B	1100
C	1110
D	1111
E	11?1

The “?” in character 3 in taxon E could be either a “0” or a “1”. If the “?” is coded as a “0” there are four resulting cladograms:



If the “?” is coded as a “1” there is only a single resulting cladograms:



We wouldn't choose an assignment that added steps to the tree, so the cladogram retained is the last one.

An unobserved state -It is also possible to image that the “?” represents a third state that has not yet been observed. In other words, the “?” may represent “0” or “1” or

“2”. This will give different results from the above when more than one taxa has a missing entry. Consider the following dataset:

Out	0000
A	1100
B	1110
C	1111
D	11?1
E	11?1

The “?” in character 3 in taxa D and E could be either a “0” or a “1”. If the “?” is coded as a “0” in both taxa, there are four resulting cladograms:

INAPPLICABLE CHARACTERS

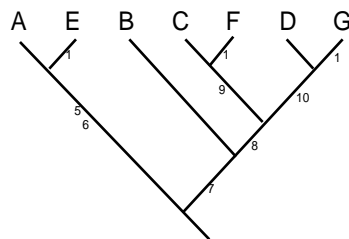
Inapplicable characters are those which are relevant for resolving one part of a cladogram, but are irrelevant, or inapplicable, to another part. For example, in an analysis of the vertebrates, having variation in the molars is relevant for grouping and separating different kinds of mammals. But the character is irrelevant for the birds, dinosaurs, lizards and snakes in the analysis because they do not have molars at all. In data matrices, many systematists use a “-” (dash) to represent inapplicable character states so as to distinguish them from missing states.

The absence of a feature may be a synapomorphy (e.g., absence of tails in the great apes is a derived feature) and should be recorded in the data matrix. The problem comes when the details of the missing feature are coded in other characters. For each of these characters we would have to record over and over again “absent” for those taxa that lack the feature. This has the effect of producing several synapomorphies for what is only a single observation: these organisms lack the

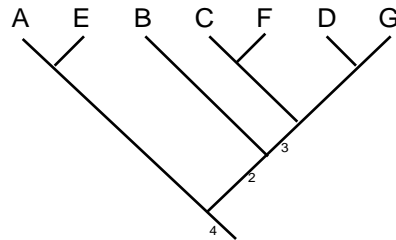
feature. This could be misleading if the absence of the feature is in any way homoplasious. Consider the following example:

Characters:	
1.	Widget: (0) present; (1) absent.
2.	Widget innervated: (0) laterally; (1) posteriorly; (2) absent.
3.	Widget ducts: (0) empty into the stomach; (1) empty into storage reservoir adjacent to the stomach; (2) absent.
4.	Widget: (0) has projection extending anterior to the base of the esophagus; (1) has projection that does not extend to the base of the esophagus; (2) absent.
5.	Some other character present in all taxa: (0) first state; (1) second state.
6.	Some other character present in all taxa: (0) first state; (1) second state.
7.	Some other character present in all taxa: (0) first state; (1) second state.
8.	Some other character present in all taxa: (0) first state; (1) second state.
9.	Some other character present in all taxa: (0) first state; (1) second state.
10.	Some other character present in all taxa: (0) first state; (1) second state.
	Out 0000 000000
	A 0001110000
	B 0101001000
	C 0111001110
	D 0111001101
	E 1222110000
	F 1222001110
	G 1222001101

Taxa E, F and G all lack the widget, therefore all characters giving details about widgets should not affect the placement of these three taxa. Just using the non-widget characters (characters 5-10) and the character for whether or not the widget is present (character 1), the tree we would expect is:

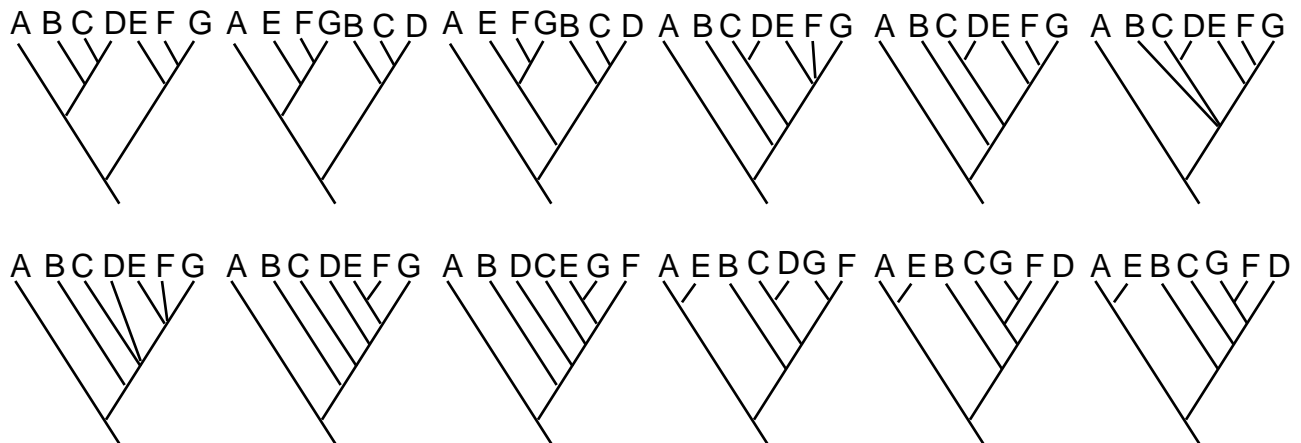


Then continuing the analysis using characters 2, 3 and 4 but without letting them affect the placement of taxa E, F, or G, the tree we expect is:



(For example: character 2's apomorphic state is in B, C and D, so that character is placed on the line leading to those three taxa; the fact that F and G don't have that state is ignored because the character does not apply to the placement of those taxa).

If the data is analyzed so absence of the widget characters in E, F, and G are treated as a state (analysis of the above data set as written, with the multistate characters analyzed unordered), 12 different trees result - none of which are the expected tree:



One solution has been to treat the inapplicable states the same as missing entries. But this too is not satisfactory. Inapplicable character states are fundamentally

different from missing entries; with inapplicable states the “-” does not represent one of the existing states observed in other taxa; for a binary character, the question mark or dash is neither a “0” or a “1”. This problem is illustrated with the following dataset:

Characters:

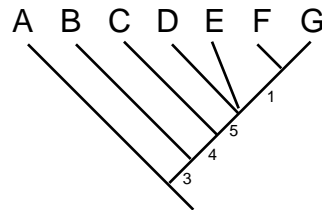
1. Thingy: (0) present; (1) absent.
2. Thingy articulated: (0) laterally; (1) posteriorly.
3. Some other character present in all taxa: (0) first state; (1) second state.
4. Some other character present in all taxa: (0) first state; (1) second state.
5. Some other character present in all taxa: (0) first state; (1) second state.

```

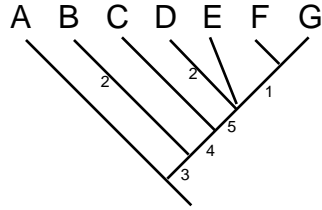
Out  00 000
A   00 000
B   01 100
C   00 110
D   01 111
E   00 111
F   1- 111
G   1- 111

```

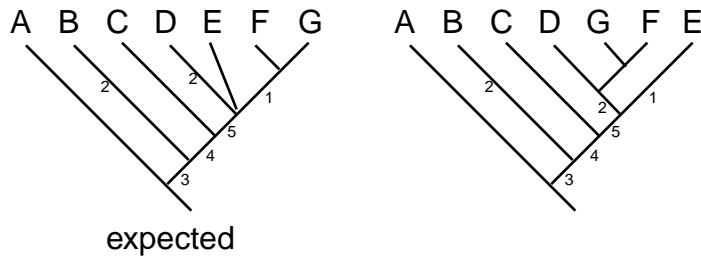
Character 2 is inapplicable for taxa F and G. The tree we expect using the other characters is:



To consider character 2 for just taxa A, B, C, D and E does not change this tree. The apomorphic state is in two taxa (B and D) but to unite those taxa would cause homoplasy in 2 other characters (4 and 5). Thus the most parsimonious solution is to conclude that character 2 is homoplasious:



But if we consider every possible assignment of “0” or “1” to the “-”, we get two trees:



The first one is the one we expect, but the second one results when G and F are resolved as having a thingy posteriorly inserted (character 2 state 1). But this is not acceptable because neither G or F even have a thingy.

There is no way to code inapplicable characters so that they will be correctly analyzed?

Meeting The Goals Of Systematics

Systematics provides a theoretical framework which gives a classification of organisms that summarizes biological information, gives insight into the evolutionary process, and is predictive. What properties should our systematic system have to meet these goals?

Stability
Informativeness
Predictability
Repeatability
Naturalness

What method of systematics would be best for achieving these?

There are many different methods for reconstructing evolution but in general they can be divided into three sorts:

1. overall similarity or phenetic methods
2. evolutionary methods
3. cladistic or parsimony methods

I am going to argue here that the best method of systematics will be phylogenetic systematics, or cladistics. Phylogenetic systematics was first described in detail by Willi Hennig the publication of his book in English in 1966 marks the beginning of a revolution in systematics and evolution. Cladistics has two major points - the evidence for relationship comes from shared derived characters (**synapomorphies**) and organisms should be grouped into monophyletic units. Hennig also proposed what he called an **auxiliary principle** - never assume that similar features of two organisms arose independently unless there is evidence from other characters to the contrary. This auxiliary principle is a **parsimony** principle. Parsimony is a principle

of logical reasoning which simply states *when two or more hypotheses exist, the one that requires the fewest assumptions is best*. In other words if one character indicated that A and B are related, but 5 characters support grouping A with C, it would be more parsimonious to group A with C.

JUSTIFYING CLADISTICS BECAUSE IT WILL RECONSTRUCT EVOLUTION BEST

Hennig argued for cladistics on the ground that synapomorphies provide the best evidence for evolution. Recently cladistics has been criticized because, when convergent evolution is common homoplasy will be misidentified as apomorphy (as is the case of long branches attracting) and cladistics will give the incorrect tree. Some systematists therefore advocate using known evolutionary trends (e.g., transitions are more common than transversions) to set probabilities and generate trees based on likelihood statistics. Most cladists reject the likelihood approach because:

1. the desire is to discover evolutionary trends from the tree, not assume them and then force the data to fit the preconceived idea (i.e., likelihood is inductive)
2. the best way to discover homoplasy is through a new analysis involving yet more characters and more taxa (after all, long branches do not exist if enough taxa are included in an analysis). In other words, phylogenetic analysis should follow the principles of hypothetico-deductive testing, rather than inductive reasoning.

REPEATABILITY - PREFERENCE OF CLADISTICS OVER TRADITIONAL (SYNTHETIC) METHODS

Synthetic and cladistic taxonomy differ from each other primarily in the repeatability of the method. Cladistics, since it follows a more precise analytical procedure for handling data, is more repeatable and therefore more scientific. Synthetics is not repeatable and, therefore, not science but learned opinion.

A synthetic taxonomist must construct elaborate assumptions about the pathways of evolution. The synthetic classification is then based on these assumptions. Rarely do two synthetic taxonomists independently make the same assumptions; even if they do, they rarely construct the same classification. From this cause, the method is not repeatable. Synthetics can also lead to circular reasoning if one attempts to use the classification to support any theory of the evolution of the group (Wagner, 1969).

Synthetic hypotheses should be more than refined stories about evolutionary pathways. The cladistic method gives the taxonomist the ability to be more objective because it allows formulation of testable hypotheses about phylogenetic relationships among organisms without elaborate presupposition stories about those relationships. This is true because cladograms are formulated from an analysis of character data without previous assumptions about evolutionary processes. The hypothesis can be tested by cladistic analysis of new character data where agreement between first and second sets of data strengthens the hypothesis and disagreement refutes it. Furthermore, such analyses are repeatable. A cladist can send his character data to any other cladist and both will construct exactly the same classification - even if the second taxonomist is unfamiliar with the taxa or the data.

INFORMATIVENESS - PREFERENCE OF CLADISTICS OVER PHENETICS

Information content and naturalness. Sneath and Sokal (1973) argue that the purpose of biological classification is not to reflect evolution, which is unknowable, but to create an indexing device through which storage and retrieval of information about organic diversity are facilitated. To fulfill this purpose they maintain that a classification should have high information content and be natural. They assert that grouping by overall similarity (phenetics) will best achieve such a classification. However, in a series of papers, Farris (1977, 1979a&b, 1980, 1982, 1983)

shows that cladistic methods will produce classifications that have more information content and naturalness than those resulting from the phenetic approach. Therefore, when judged by the phenetic criteria, cladistic classifications are superior to phenetic ones.

Information content was discussed and defined by Mill (1874, p466):

"The ends of scientific classification are best answered when objects are formed into groups respecting which greater number of general proposition more important, than could be made respecting any other groups into which the same things could be distributed."

Simply put, information content of a classification is how much information it conveys about character state distribution. Information is contained in a classification when the branching pattern of the tree describes the distribution of the characters. For example, consider the following two trees for three taxa **A**, **B**, and **C**, their common ancestor **Anc**, and the data for two characters:

	Characters			
Taxa	I	II		
A	1	1	<p style="text-align: center;"><i>a</i></p>	<p style="text-align: center;"><i>b</i></p>
B	1	1		
C	0	0		
Anc	0	0		

In tree *a*, taxa **A** and **B** are sister taxa, and **C** is the sister taxon to the **AB** clade. In tree *b*, taxa **C** and **B** are sister taxa, and **A** is the sister taxon to the **BC** clade. Both of the characters change from state 0 to state 1 in **A** and **B**, indicating that **A** and **B** should be placed together. Consequently, the branching pattern of tree *a* describes the changes in the character data perfectly. Tree *b* branches in a pattern that is opposite to the changes in the characters. Thus, tree *a* has higher information content than tree *b*.

Information content can be indicated simply by counting the number of

deviations of the character state distribution from the tree. The lower the number of deviations, the higher the information content (Shuh and Polhemus, 1981). In tree *a*, there are no deviations. In tree *b*, there are two convergent characters (state 1 appears to have arisen twice - once in taxon A and once in taxon B).

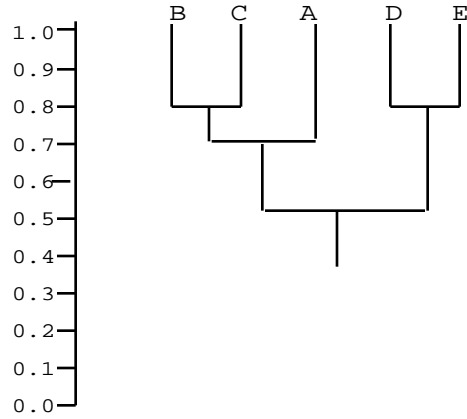
Farris (1977, 1979a&b, 1980, 1982, 1983) has argued that parsimoniously grouping by synapomorphy (shared derived character states) rather than similarity (phenetics) will always give a more informative classification. This can be demonstrated by examining a phenogram and a cladogram for the following data set of 10 characters for five species (A-E):

Taxa	Characters									
	1	2	3	4	5	6	7	8	9	10
A	+	+	+	+	+	+	+	+	+	-
B	+	+	+	-	-	+	+	+	-	-
C	+	+	+	+	-	+	+	+	-	+
D	+	+	-	-	-	+	-	-	-	-
E	+	+	-	-	-	-	+	-	-	-

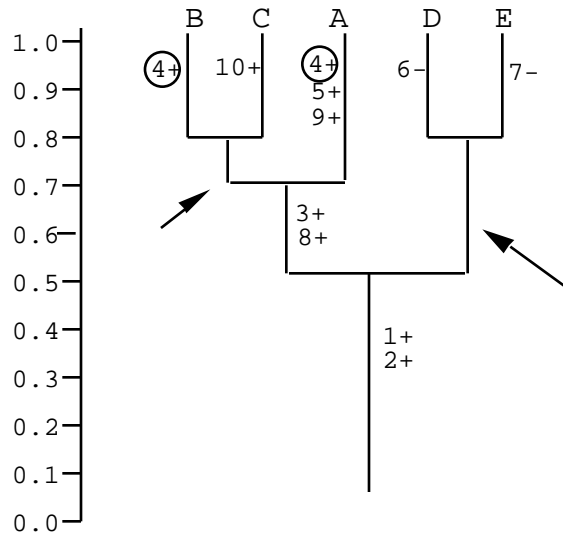
The similarity is:

	A	B	C	D	E
A	-				
B	70%	-			
C	70%	80%	-		
D	40%	70%	50%	-	
E	40%	70%	50%	80%	-

The resulting relationships based on similarity are:



What information does the branching diagram convey about the characters:



In the phenogram, character 4 shows a convergence in taxa B and A. Furthermore, two of the branches - the BC group and the DE group are not defined by any characters (arrows in above phenogram).

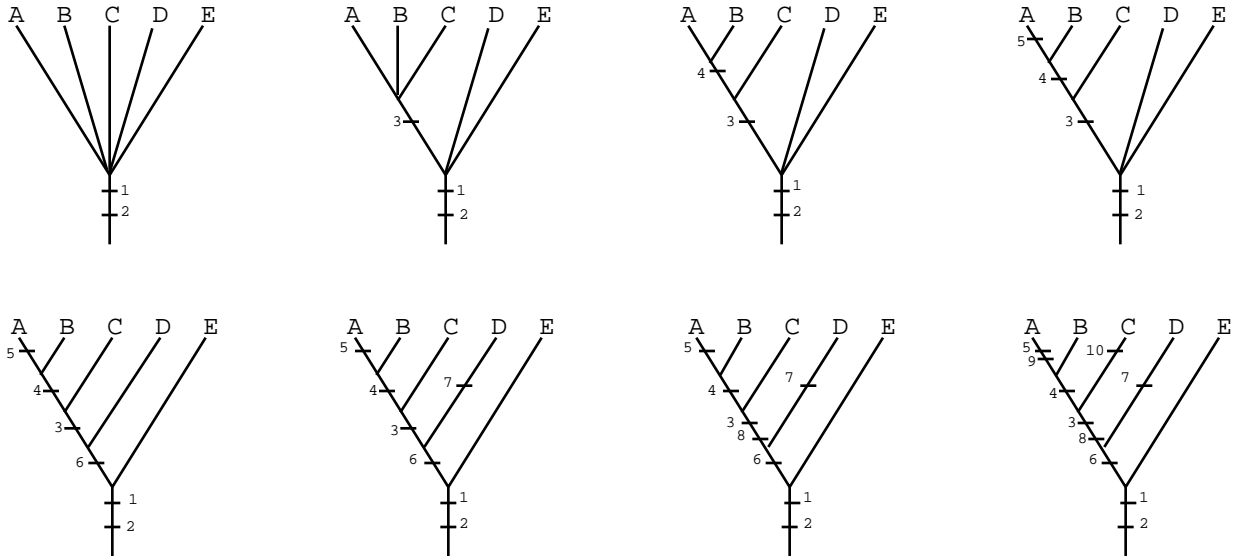
Now suppose that an outgroup is also examined, and the data set is rewritten so that the plesiomorphic state (primitive state) is represented with a 0 and the apomorphic state (advanced state) is represented with a 1. This character polarity is determined using the outgroup method:

Taxa	Characters									
	1	2	3	4	5	6	7	8	9	10
A	+	+	+	+	+	+	+	+	+	-
B	+	+	+	-	-	+	+	+	-	-

C	+	+	+	+	-	+	+	+	-	+
D	+	+	-	-	-	+	-	-	-	-
E	+	+	-	-	-	-	+	-	-	-
OUT	-	-	-	-	-	-	+	-	-	-

Taxa	Characters									
	1	2	3	4	5	6	7	8	9	10
A	1	1	1	1	1	1	0	1	1	0
B	1	1	1	0	0	1	0	1	0	0
C	1	1	1	1	0	1	0	1	0	1
D	1	1	0	0	0	1	1	0	0	0
E	1	1	0	0	0	0	0	0	0	0
OUT	0	0	0	0	0	0	0	0	0	0

The resulting cladogram is:



In the cladogram, none of the characters show convergence or reversals.

Furthermore, all of the branches are defined by one or more character states.

The branching pattern of the cladogram deviates from the character state distributions less than that shown by the phenogram. The cladogram, therefore, describes more of the character state changes and has a higher information content.

Farris has demonstrated that this will always be the case because the cladistic method parsimoniously groups by advanced character states in a hierarchical

branching sequence, whereas phenetics averages all of the character data over the entire tree. There are two points to this argument:

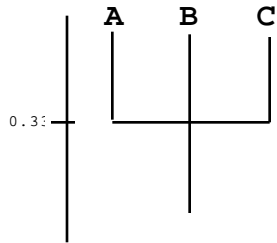
First, the classification that allows the data to be described in the smallest number of character state changes corresponds to the most parsimonious tree. It is also the classification with the highest information content. It is obvious that trees that correspond best to the branching character state distribution have the fewest steps, that is the most parsimonious. Therefore, of any two or more classifications, the one which is most parsimonious is the one with the greater information content. Since cladistics seeks the most parsimonious tree and phenetic methods do not, cladograms should always have higher (or at least equal is the cladogram and phenogram are identical) information content than a phenogram.

The second point is more complex. Taxonomists have long been aware that the branching pattern of classification will rarely reflect the distribution of all of the character states. Usually, several characters correspond closely but not exactly in the distribution of their states. The result is that the classification most informative about any one of those characters will differ slightly from those most informative about any other character. Pheneticists have generally maintained that this problem is best solved by averaging the data (overall similarity). This will create a classification in which every cluster of taxa is informative about every character. If the distributions of the states of two characters differ, a compromise tree with a branching pattern intermediate between that dictated by the two characters is produced. The result is the loss of information content. A simple example will illustrate:

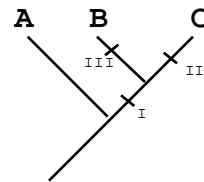
Taxa	Characters		
	I	II	III
A	0	0	0
B	1	0	1
C	1	1	0

Simple Matching Similarity Matrix

	A	B	C
A	-		
B	0.33	-	
C	0.33	0.33	-



Phenogram



Cladogram

The three characters indicate three different relationships: Taxa **B** and **C** would be united if only character I is consulted; Taxa **A** and **B** would be united if character II was considered because they have the same state; Character III is the same in taxa **A** and **C**. The phenogram calculated from these data shows a branching pattern that is intermediate between those indicated by the three characters. But the phenogram has no information content because it does not reflect any of the character state distributions.

The compromise is not necessary when one creates classifications that are hierarchical like the natural hierarchy created by descent with modification. Some groups in a hierarchy are subsets (or supersets) of other groups in the same classification. A simple consequence is that it is not possible for every group to be informative on every character considered (Farris, 1980). For example, one character may describe relationships among genera being classified; another character may describe the species of one of the genera. Unlike phenetics, the cladistic method describes the character state changes topographically on the tree because it joins taxa by actual shared character states. The taxa corresponding to characters that have

different distributions are incorporated into a single classification, with the different distributions being placed at different branch points. The cladogram shown above describes all of the character state distributions perfectly.

Farris has demonstrated that naturalness is equivalent to information content and that maximizing one will maximize the other. A natural classification, as defined by Whewell (1840, p. 512) is:

"... that arrangement obtained when one set of characters coincides with the arrangement obtained from another set."

The natural classification of a group is the one supported by the greatest agreement of the character state distributions. Since the classification with the highest information content will have the maximum congruence of characters, it will also be the most natural. Because cladograms have the highest information content, they will be the most natural.

References

- Adams, E. N., III. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology* 21:390-397.
- Bremer, K. 1990. Combinable component consensus. *Cladistics* 6:369-372.
- Camin, J. H. and R. R. Sokal. 1965. A method for deducing branching sequences in phylogeny. *Evolution* 19:311-326.
- Farris, J. S. 1969. A successive approximations approach to character weighting. *Systematic Zoology* 18:374-385.
- Farris, J. S. 1970. Methods for computing Wagner trees. *Systematic Zoology* 19:83-92.
- Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. *American Naturalist* 106:645-668.
- Farris, J. S. 1977. Phylogenetic analysis under Dollo's Law. *Systematic Zoology* 26:77-88.
- Farris, J. S. 1982. Outgroups and parsimony. *Systematic Zoology* 31:328-334.
- Farris, J. S. 1988. Hennig86, version 1.5. Distributed by the author, Port Jefferson Station, N. Y.
- Farris, J. S. 1989a. The retention index and homoplasy excess. *Systematic Zoology* 38:406-407.
- Farris, J. S. 1989b. The retention index and the rescaled consistency index. *Cladistics* 5:417-419.
- Felsenstein, J. 1978a. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology* 27:401-410.
- Felsenstein, J. 1978b. The number of evolutionary trees. *Systematic Zoology* 27:27-33.
- Felsenstein, J. 1984. The statistical approach to inferring phylogeny and what it tells us about parsimony and compatibility. Pages 169-191 in T.
- Fitch, W. M. 1971. Toward defining the course of evolution: minimal change for a specific tree topology. *Systematic Zoology* 20:406-416.

Hendy, M. D. and D. Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* 59:277-290.

Hennig, W. 1966. *Phylogenetic Systematics*. (University of Illinois Press: Urbana, Illinois).

Kluge, A. G. and J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Zoology* 18:1-32.

Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology* 40:315-328.

Maddison, W. P., M. J. Donoghue, and D. R. Maddison. 1984. Outgroup analysis and parsimony. *Systematic Zoology* 33:83-103.

Miyamoto, M. M. 1985. Consensus cladograms and general classifications. *Cladistics* 1:186- 189.

Page, R. D. M. 1989. Comments on component- compatibility in historical biogeography. *Cladistics* 5:167-182.

Swofford, D. L. 1991. When are phylogeny estimates from morphological and molecular data incongruent? Pages 295-333 in M. M. Miyamoto and J. Cracraft (ed.), *Phylogenetic Analysis of DNA Sequences* (Oxford University Press: New York, N. Y.).

Swofford, D. L. and G. J. Olsen. 1990. Phylogeny reconstruction. Pages 411-501 in D. M. Hillis and C. Moritz (ed.), *Molecular Systematics* (Sinauer Associates: Sunderland, Massachusetts).

Wiley, E. O. 1981. *Phylogenetics. The Theory and Practice of Phylogenetic Systematics*. (Wiley and Sons: New York).

Wiley, E. O., D. Siegel-Causey, D. R. Brooks, and V. A. Funk. 1991. *The Compleat Cladist. A Primer of Phylogenetic Procedures*. (University of Kansas Museum of Natural History Special Publ. No. 19: Lawrence, Kansas).