Distinguishing Phylogeny vs. Taxonomy

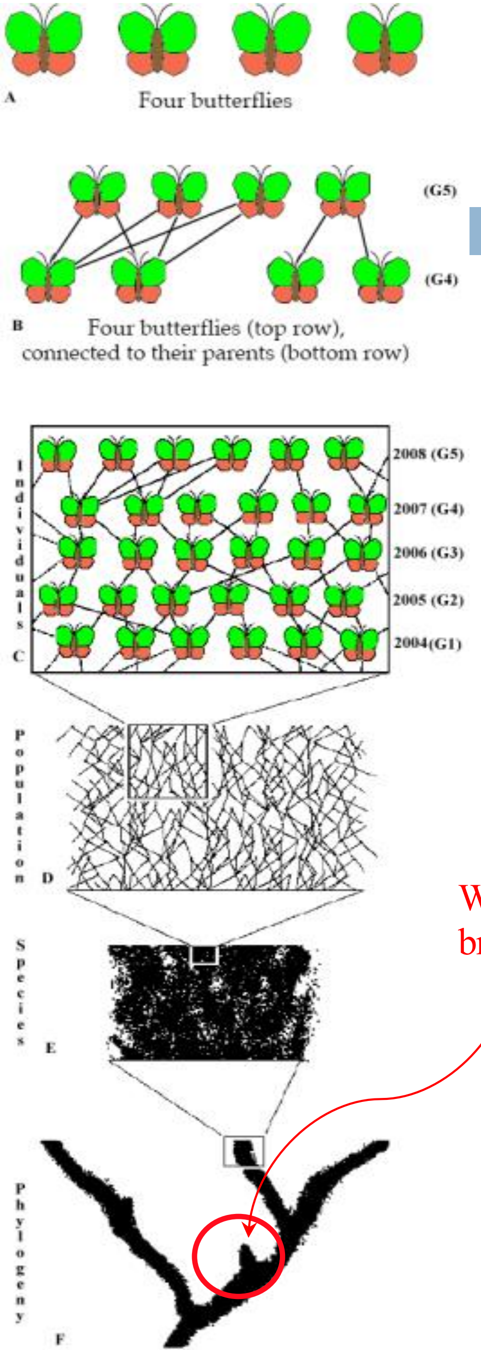
- Taxonomy – a hierarchical structure of grouping & naming or classifying organisms based on similarities and differences

- Phylogeny – Depicts the evolutionary relationships of a group of organisms.

Phylogenetic trees indicate when/where species diverged from a common ancestor.

Phylogenetic trees show lines of descent

The branching patterns that make up a phylogenetic tree are not usually observed directly, rather they are inferred from data



What does this dead-end branch depict?

Image adapted from: Baum, D.A. and Smith S.D. Tree thinking: an introduction to phylogenetic biology. Roberts,

Before we can create phylogenetic trees, we need to ask…

How are organisms classified?

- Let's start with a set of species named as follows…
  - A
  - B
  - C
  - D
  - E
  - F

- We have evidence that:
  - B and C are more closely related to each other than to any other species
  - A is related to B and C, but not as closely as B and C are related to each other
  - D, E and F are more closely related to each other than to A, B or C
- Let's start here and learn a simple way to organize the species based on relatedness.

How are organisms classified?

We know:

- B and C are closely related

- Start by grouping the most closely related organisms together.
- Use parentheses to put them together in a set…
- …and a comma to show that they share a common ancestor.

B C

(B C)

(B,C)

How are organisms classified?

We know:

- B and C are closely related
- A is related to both B and C, but not as closely as B and C are related to each other

- Now add the next most closely related organism
- Place this organism next to the first set,
- add parentheses to group them into a larger set,
- and a comma to show the common ancestor.

A (B,C)　　(A (B,C))　　(A,(B,C))

How are organisms classified?

- What about the other organisms?
  - D
  - E
  - F

  - D, E and F are more closely related to each other than to A, B or C.

- If D, E and F are equally related to each other (or you're not sure!) put them together in a set using parentheses.
- And separate them with commas to show they share a common ancestry.

D E F

(D E F)

(D,E,F)

How are organisms classified?

- Let’s bring this all together.
- Because the DEF set is more distantly related from the ABC set, we keep the 2 sets separate but put them next to each other.

(A,(B,C)) (D,E,F)

- Then we add a final set of parentheses to enclose all of the species as one larger set.
- Don’t forget the comma, to show the common ancestor between the two sets that make up the larger set.

((A,(B,C)),(D,E,F))

How are organisms classified?

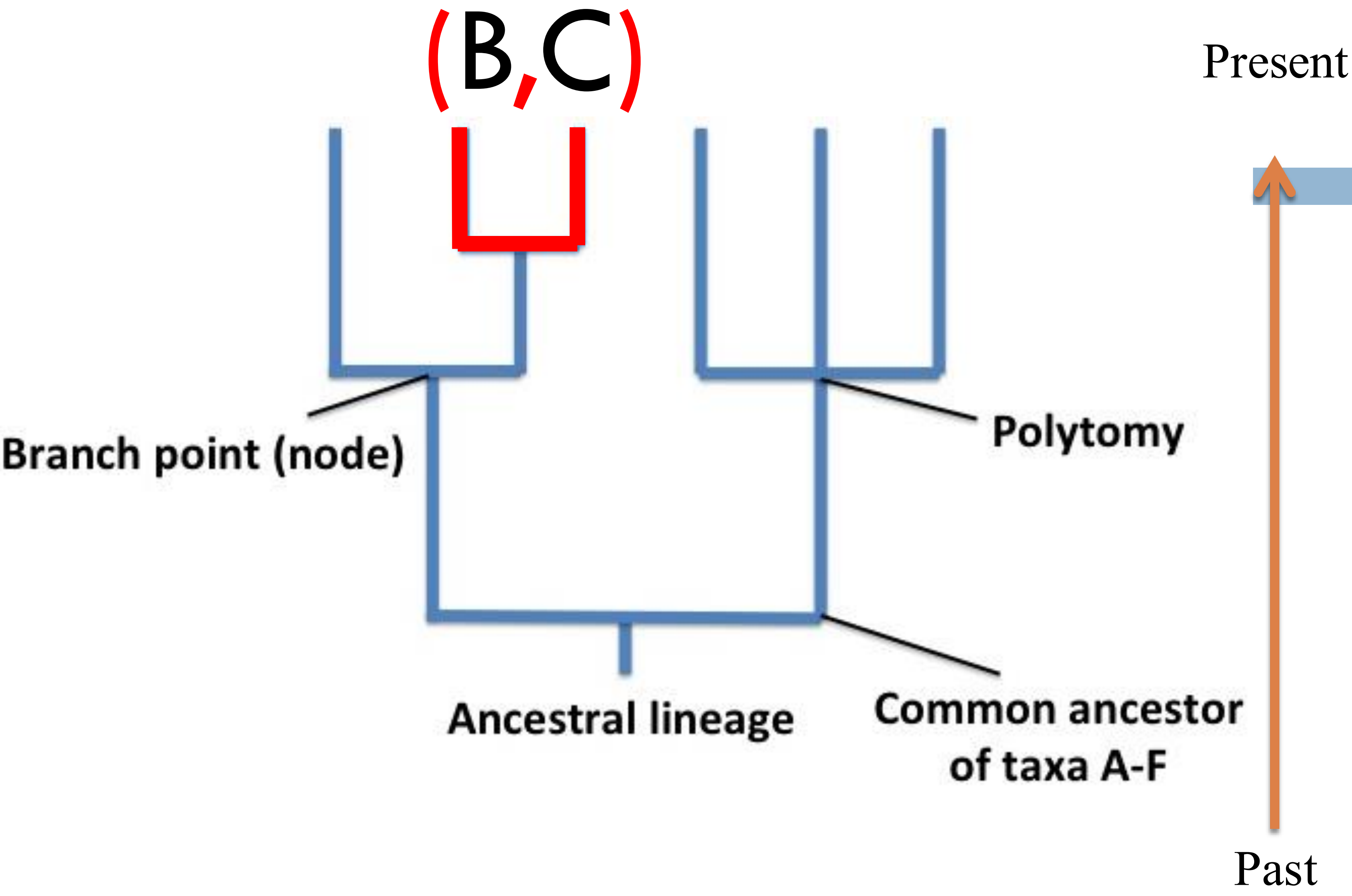
- One last thing!
- To show that you have completed the tree, you add a semi colon to the end of the set.

((A,(B,C)),(D,E,F));

How are organisms classified?

((A,(B,C)),(D,E,F));

- *This is a phylogenetic tree*



(B,C)

Present

Branch point (node)

Polytomy

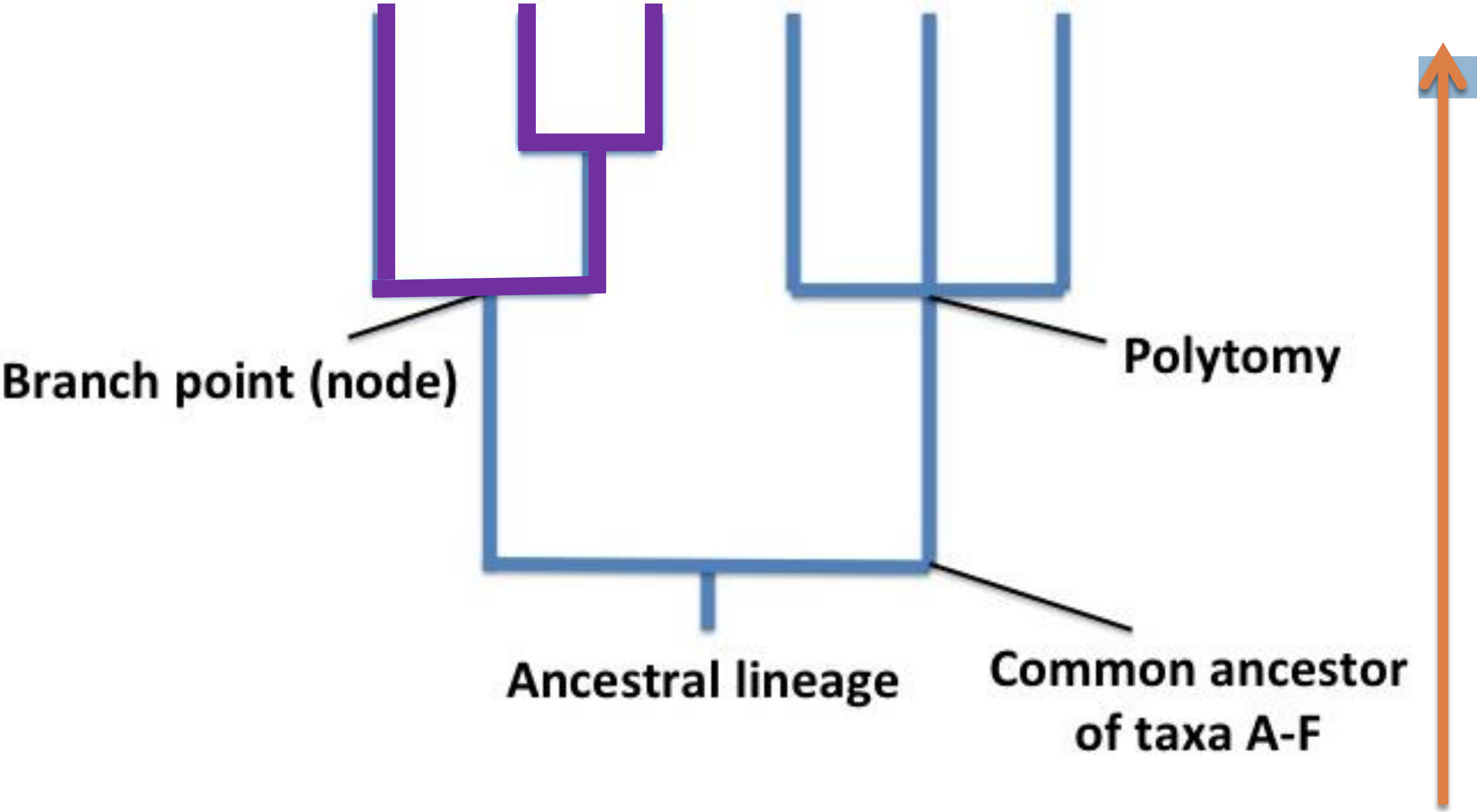
Ancestral lineage

Common ancestor of taxa A-F

Past

(A,(B,C))

Present



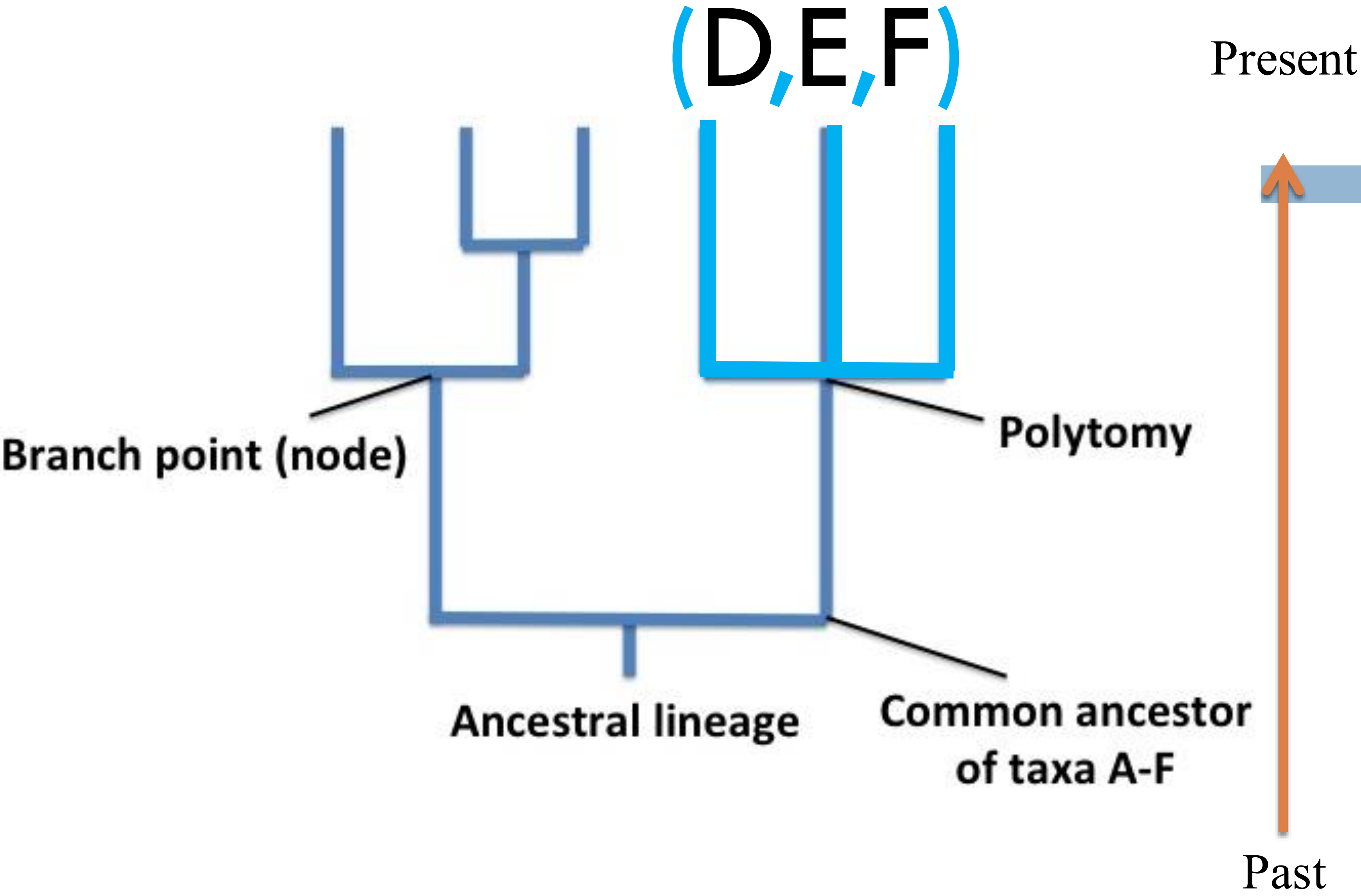
Branch point (node)

Polytomy

Ancestral lineage

Common ancestor
of taxa A-F

Past



(D,E,F)

Present

Branch point (node)

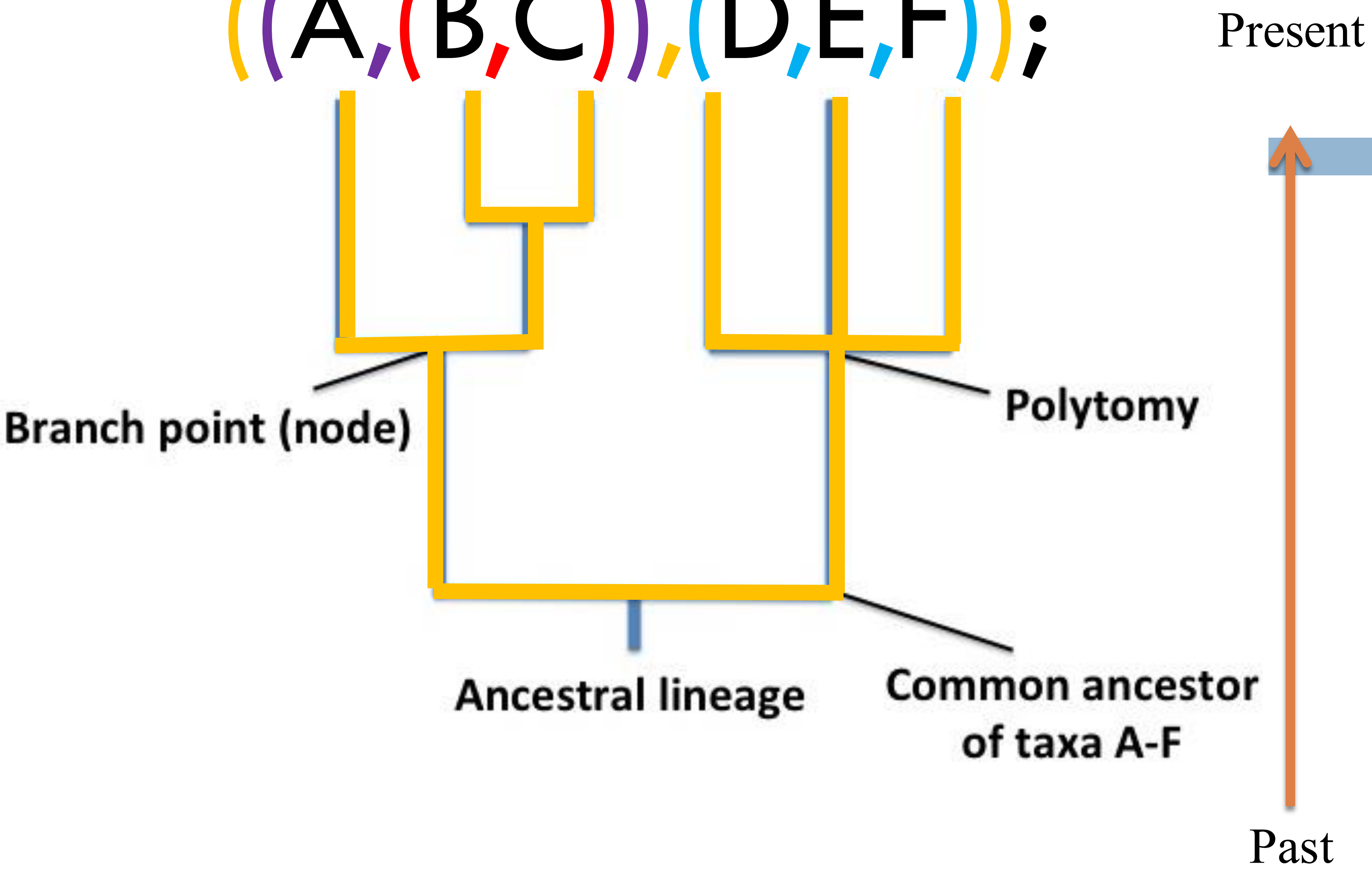
Polytomy

Ancestral lineage

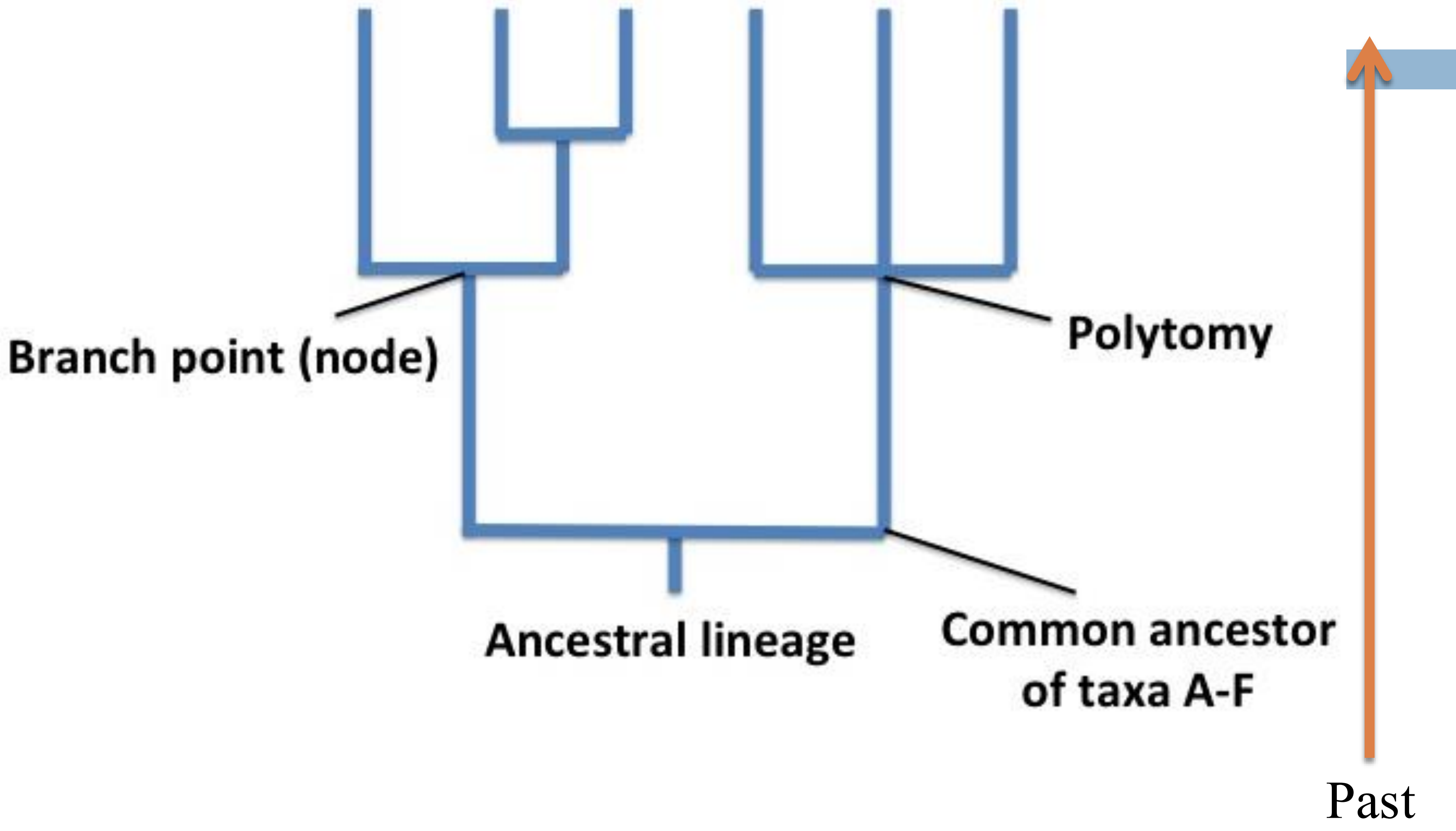
Common ancestor of taxa A-F

Past

((A,(B,C)),(D,E,F));


Present
Polytomy
Branch point (node)
Ancestral lineage
Common ancestor of taxa A-F
Past

((A,(B,C)),(D,E,F));



Present

Branch point (node)

Polytomy

Ancestral lineage

Common ancestor of taxa A-F

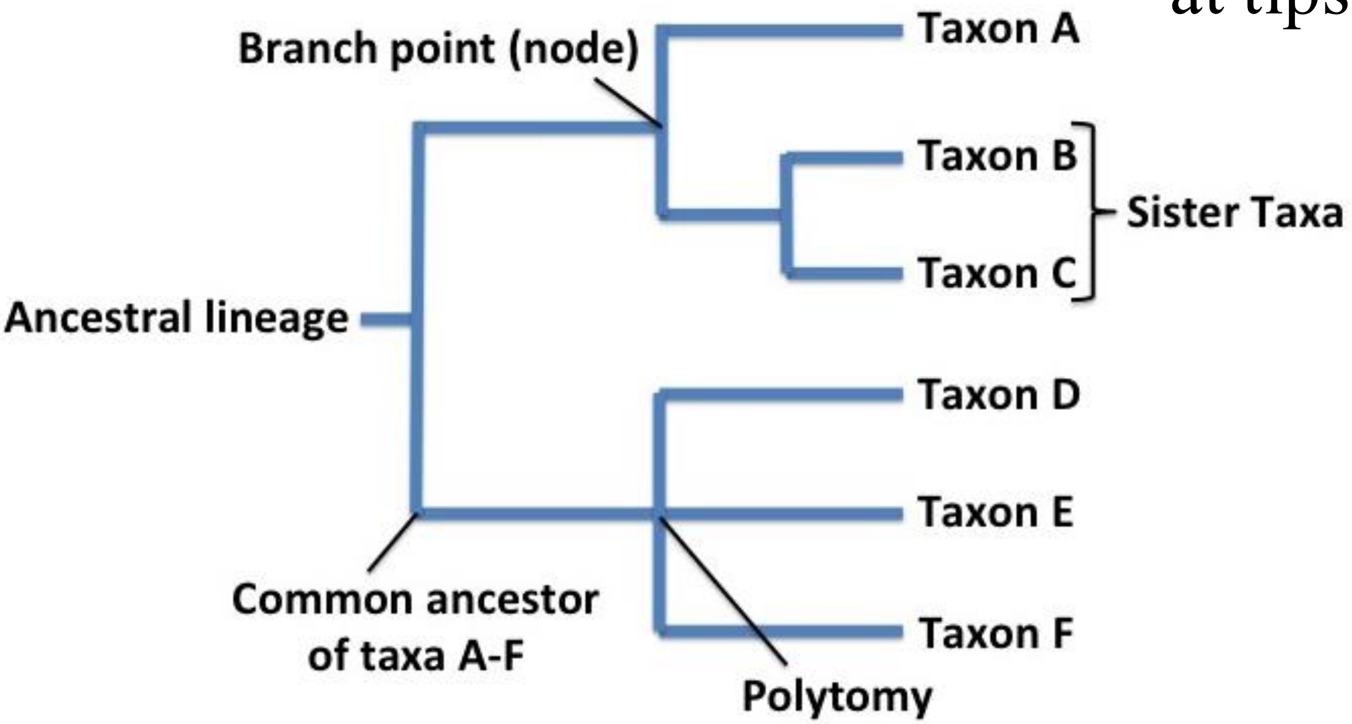
Past

Some common terminology

Nodes indicate common ancestor

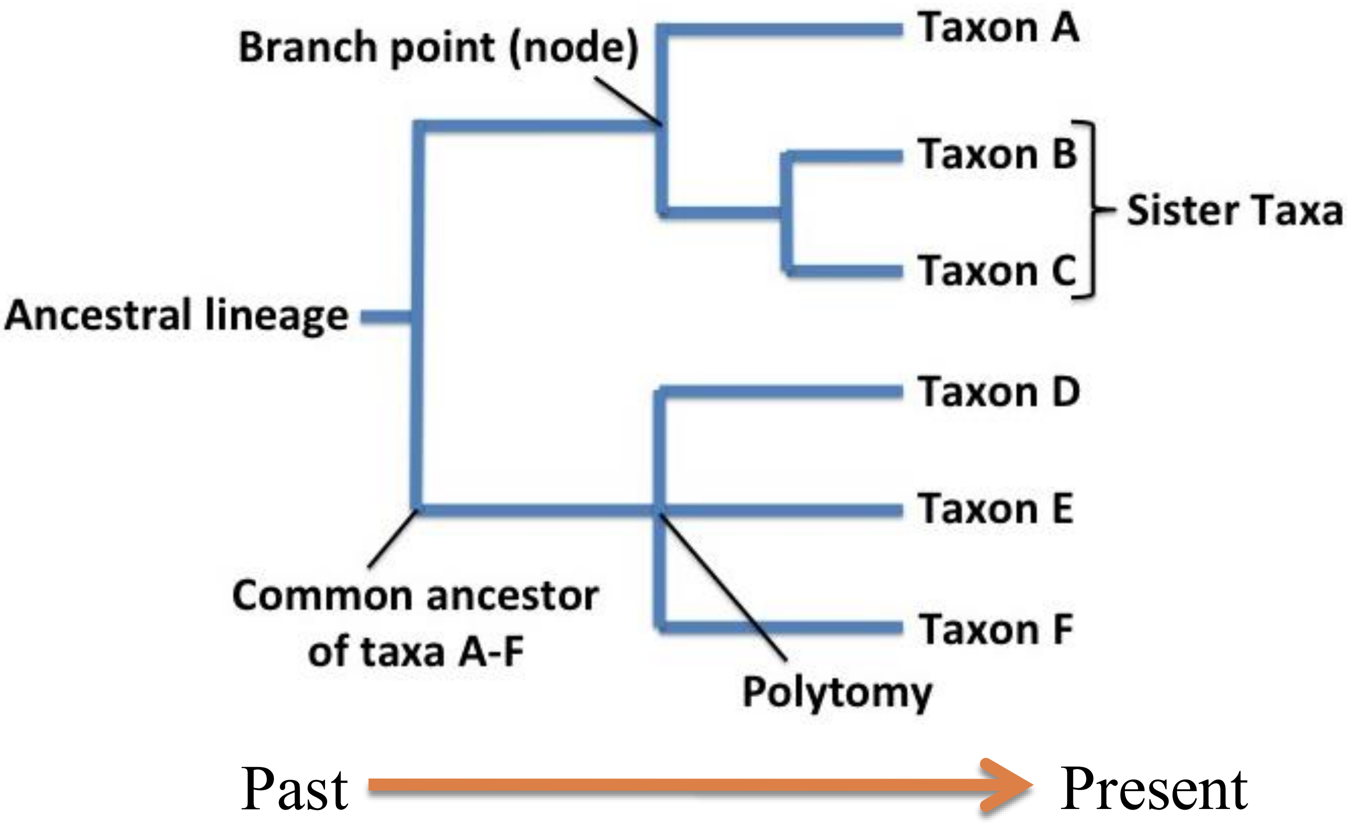
Branches indicate evolutionary path

Descendants (Current taxa) at tips

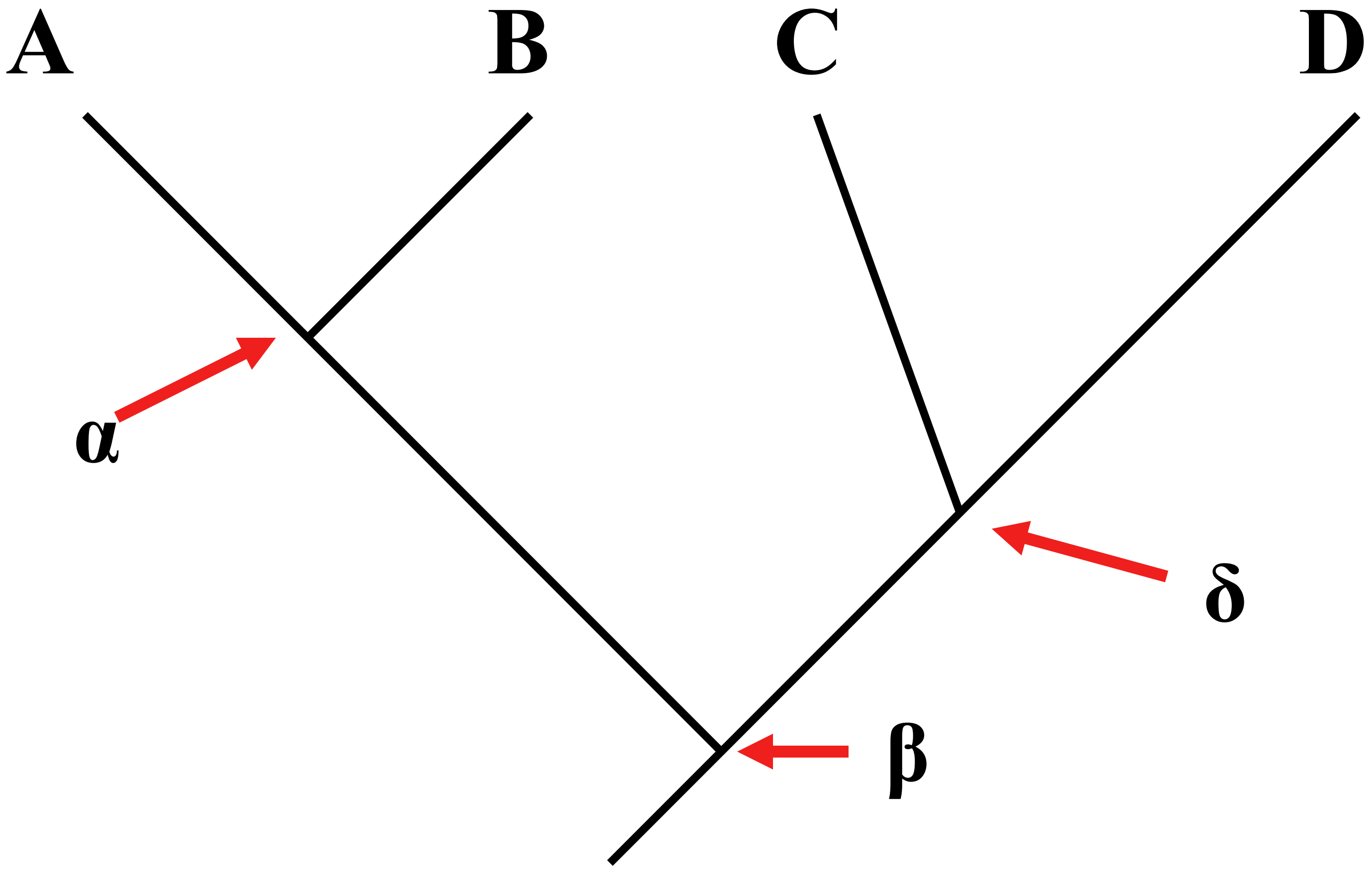


Past → Present

Now that we've seen how trees are put together let's look at how to interpret them.



Unless stated otherwise, assume that absolute time is NOT represented

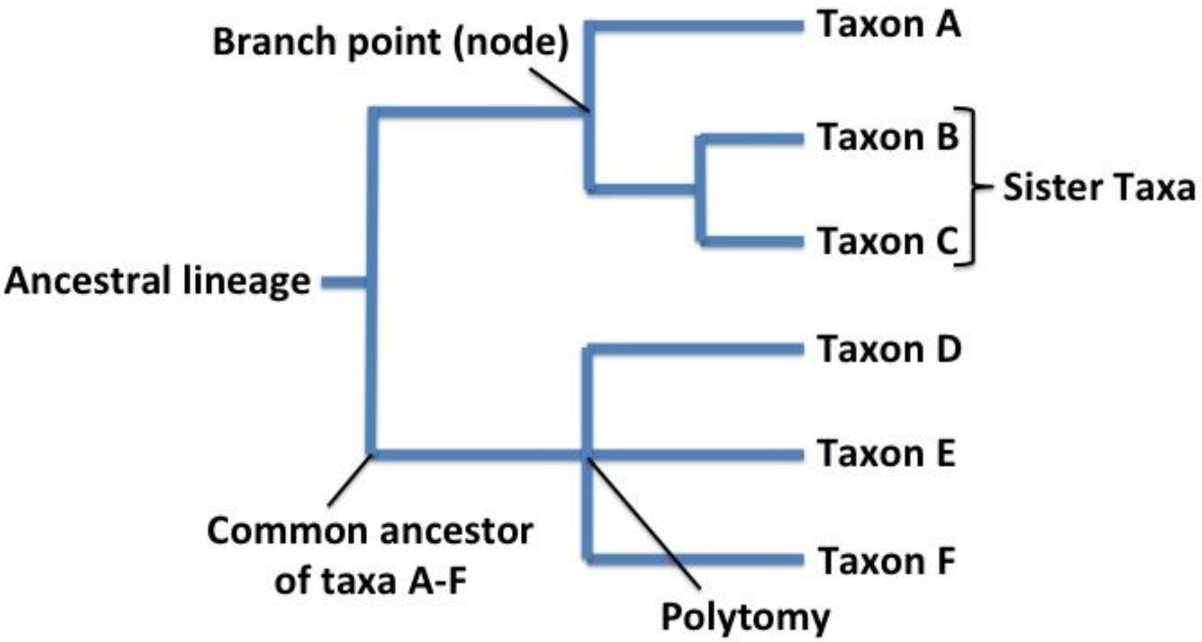


Was α before or after β?

Was α before or after δ?

More terminology

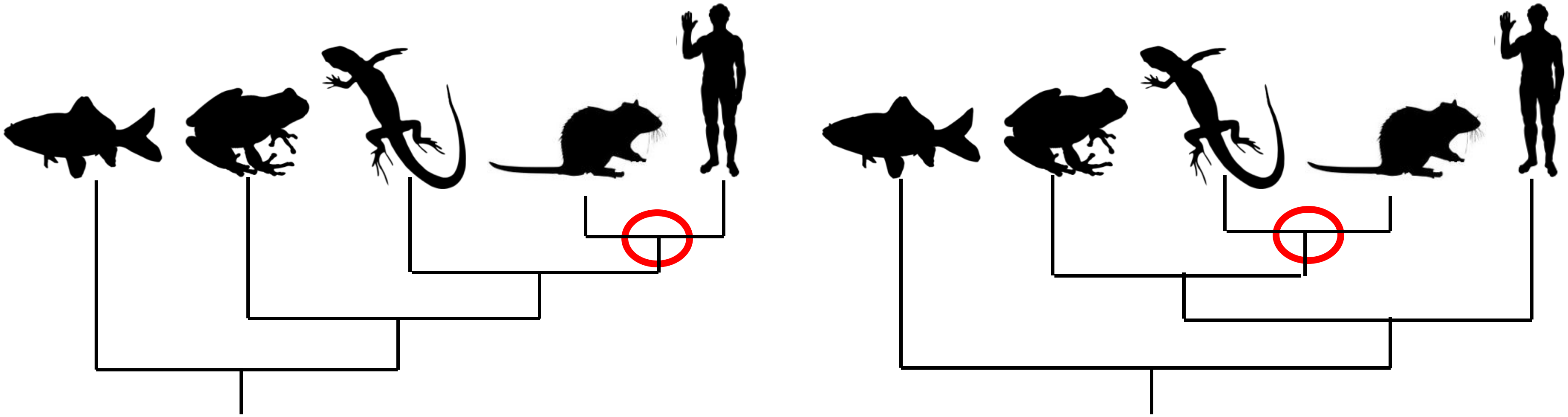
A cladogram indicates relatedness *only!*



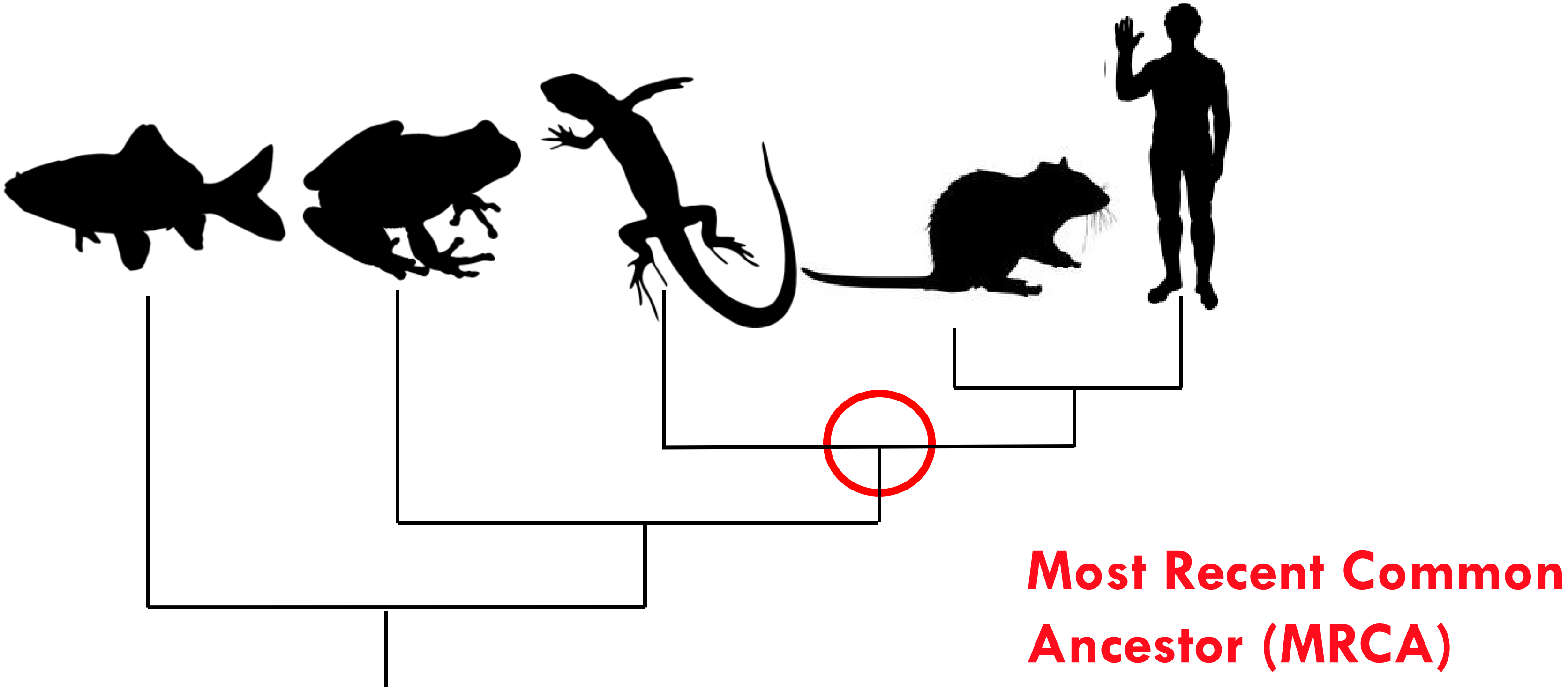
Sister taxa are most closely related because the MRCA (**m**ost **r**ecent **c**ommon **a**ncestor) that they share is more recent than the ones they share with another taxon on this tree

Branch length is not always meaningful

Same order – different tree?



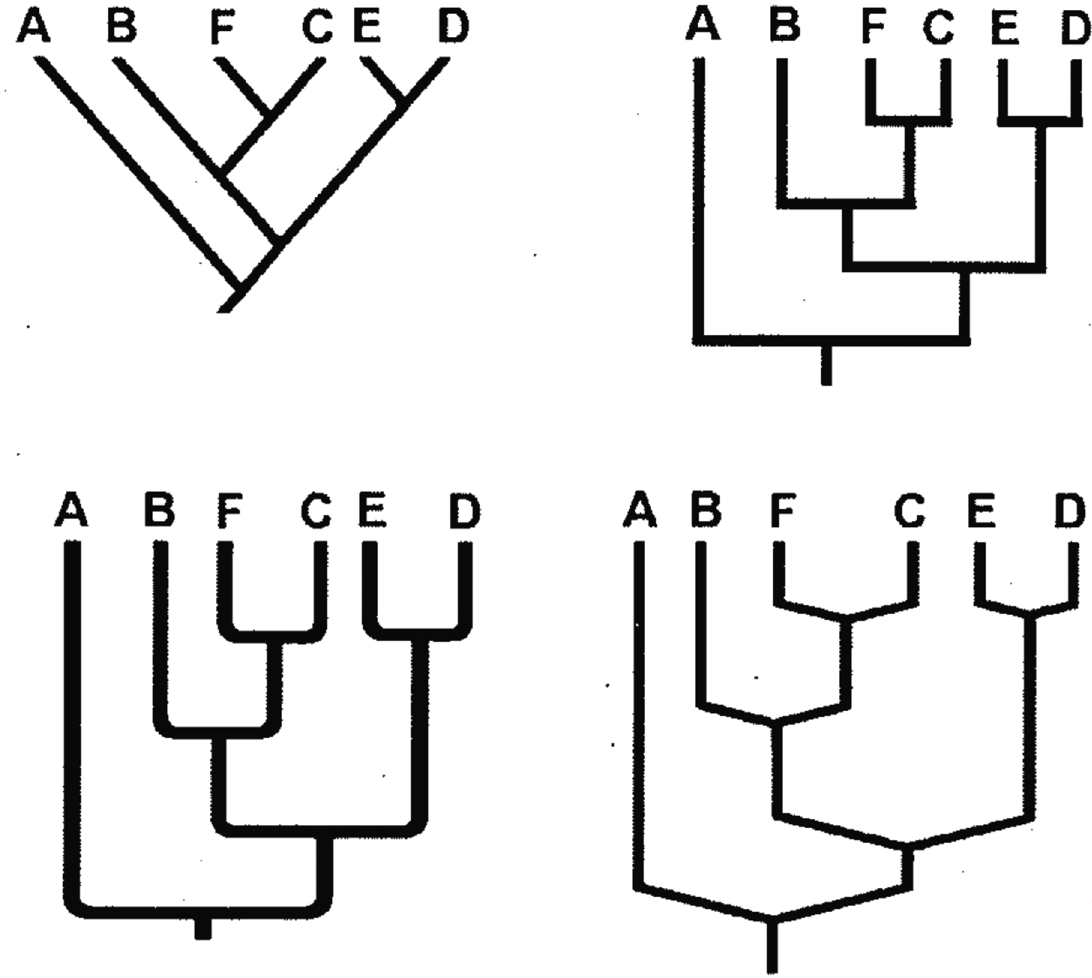
Don’t be distracted by morphological similarity


Most Recent Common Ancestor (MRCA)

Is a lizard more closely related to a frog or a human?

Images via Wikimedia Commons used under Creative Commons Attribution 2.0 Generic License

Trees can be drawn in various ways yet contain the same phylogenetic information

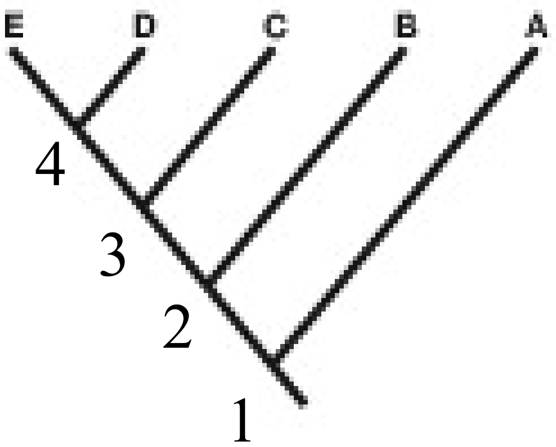


Phylogenetically, these trees are identical.

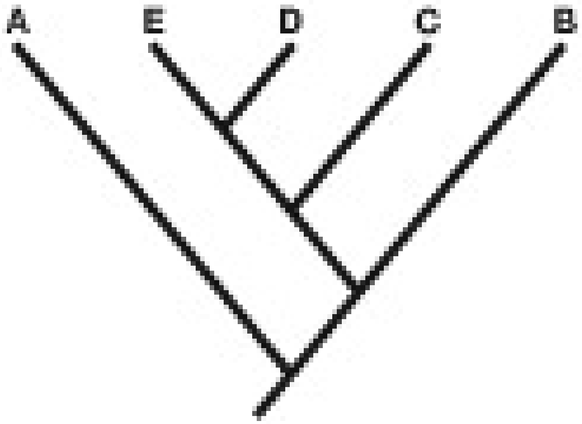
Gregory 2008. Evo Edu Outreach 1:121-127

Rotation around the nodes does not change the information in a tree.

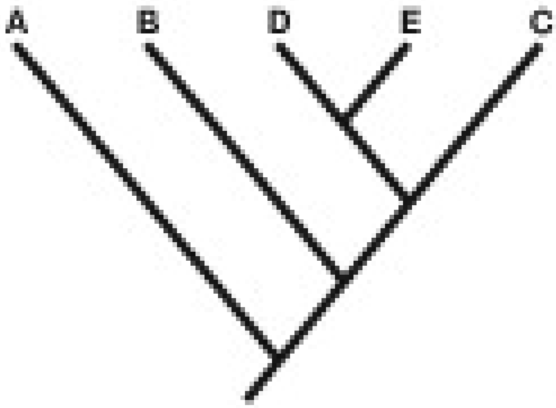
These three trees are identical



Tree 1



Tree 2



Tree 3

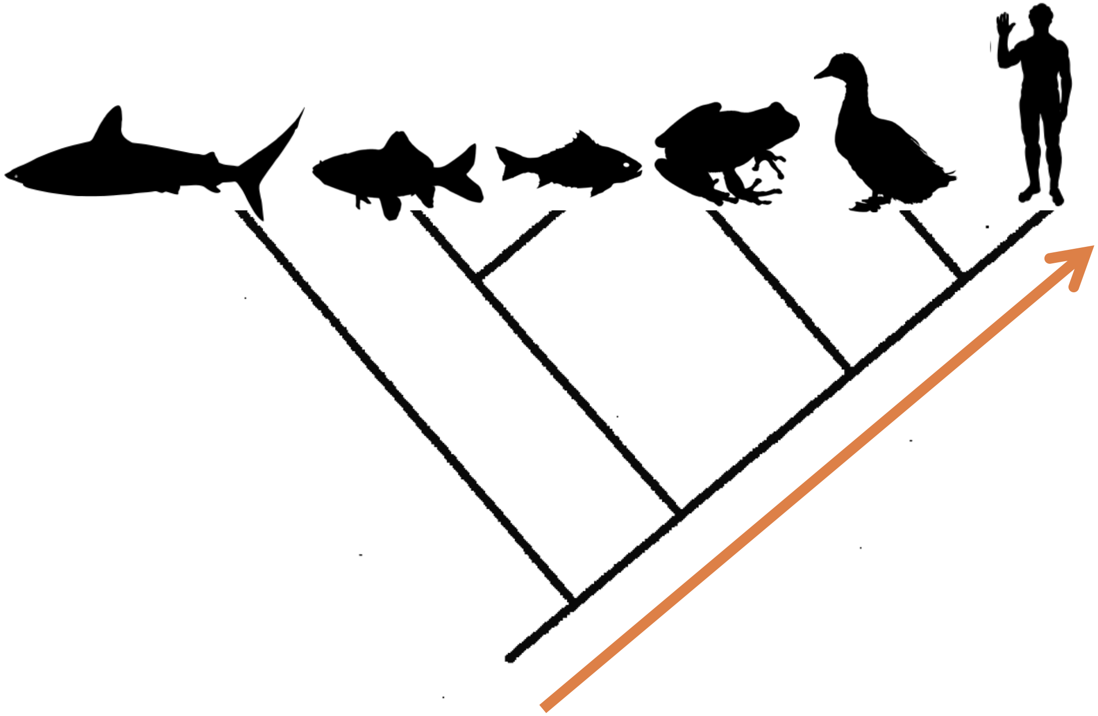
If we rotate around the first node…

If we rotate around the 1st, 2nd, 4th nodes…

…we get this

…we get this

There are no "main lines" in evolution



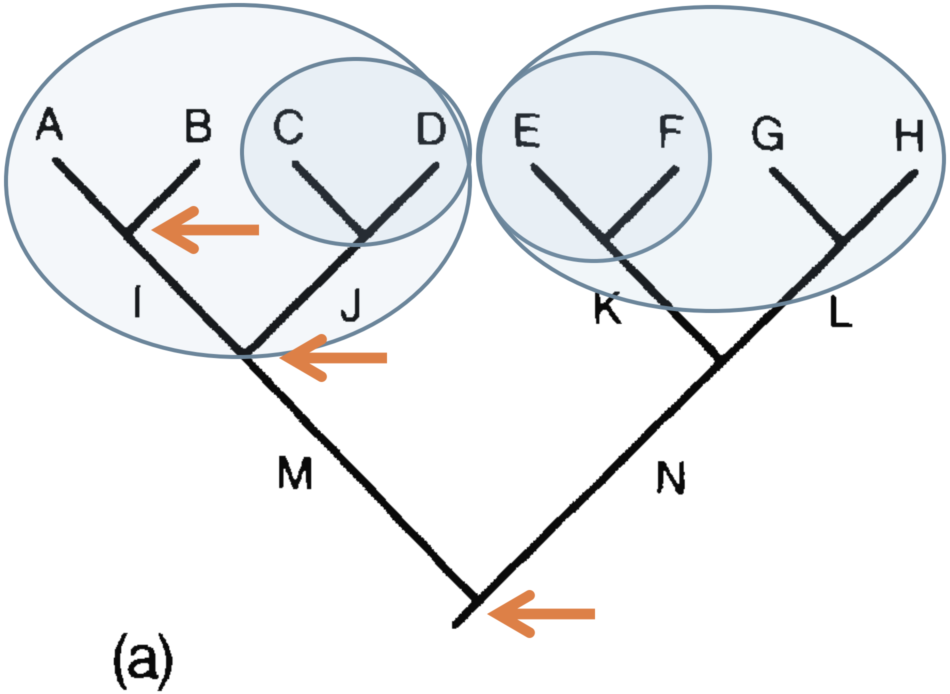
Just because the human has a "straight shot" does not mean it is more related to the common ancestor than the others.

The "straight shot" also does not imply that the human (or the shark) is more advanced than the other taxa.

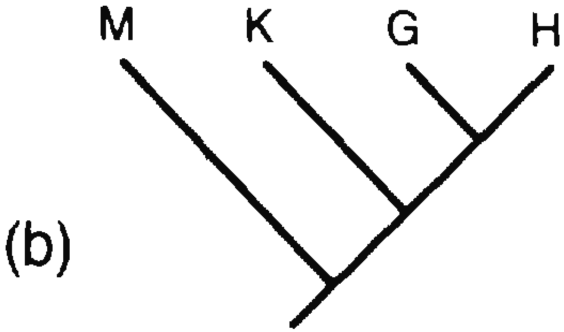
Images via Wikimedia Commons used under Creative Commons Attribution 2.0 Generic License

Clades can be expanded or collapsed

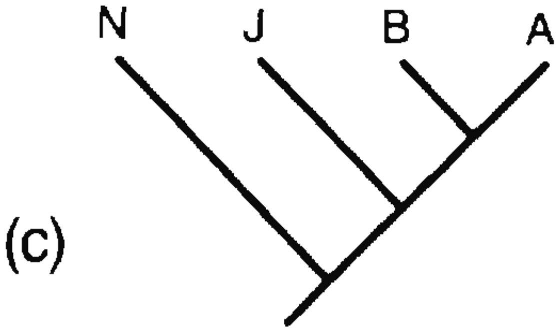
If we collapse these two regions…



…we get this tree

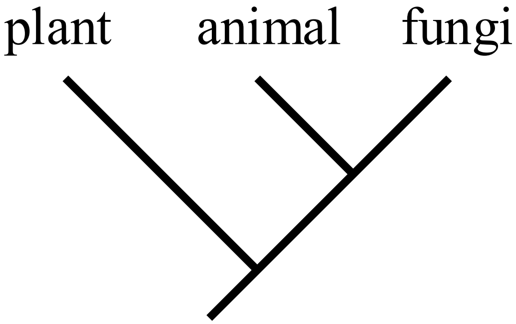


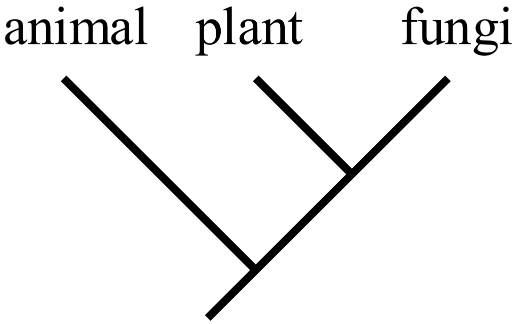
…we get this tree

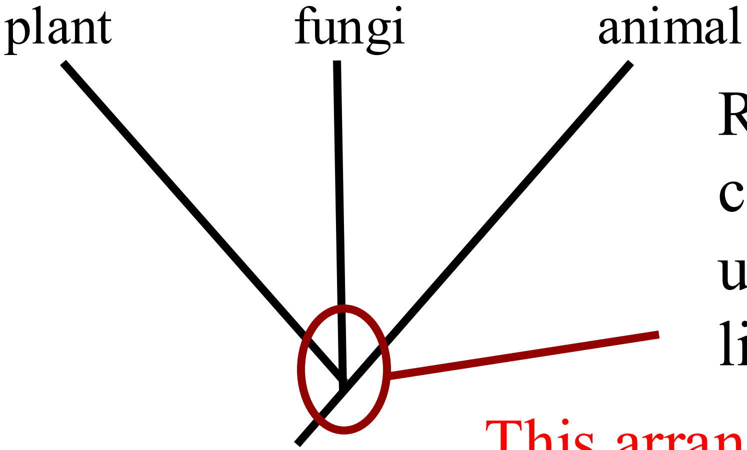


A tree can depict polytomy

If we are unsure which of these is most likely:





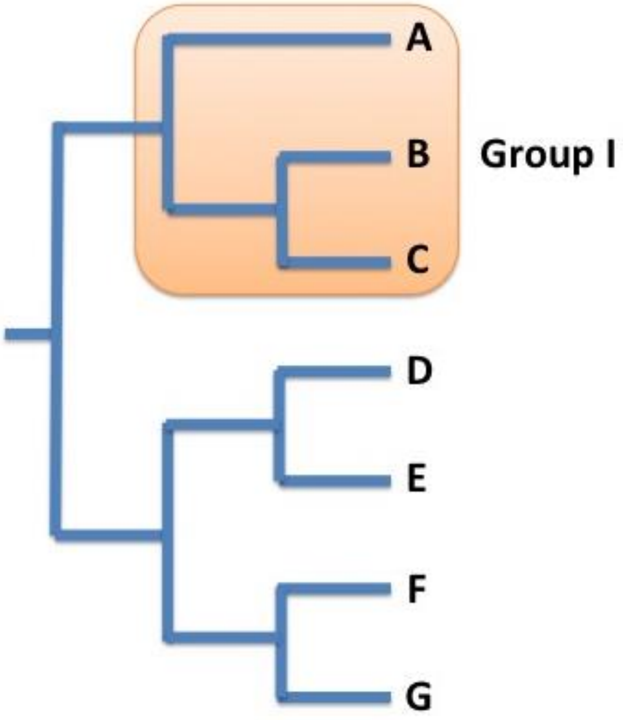


Rearranging them to meet at a common point shows uncertainty about how lineages relate to each other.

This arrangement is a called a polytomy.

Monophyletic groups (clades)

Monophyletic group (clade)



A clade is a grouping that includes a common ancestor and all the descendants (living and extinct) of that ancestor

Scientists try and make sure that each taxon (a named group of organisms) corresponds to a clade

How can we reconstruct phylogenies?

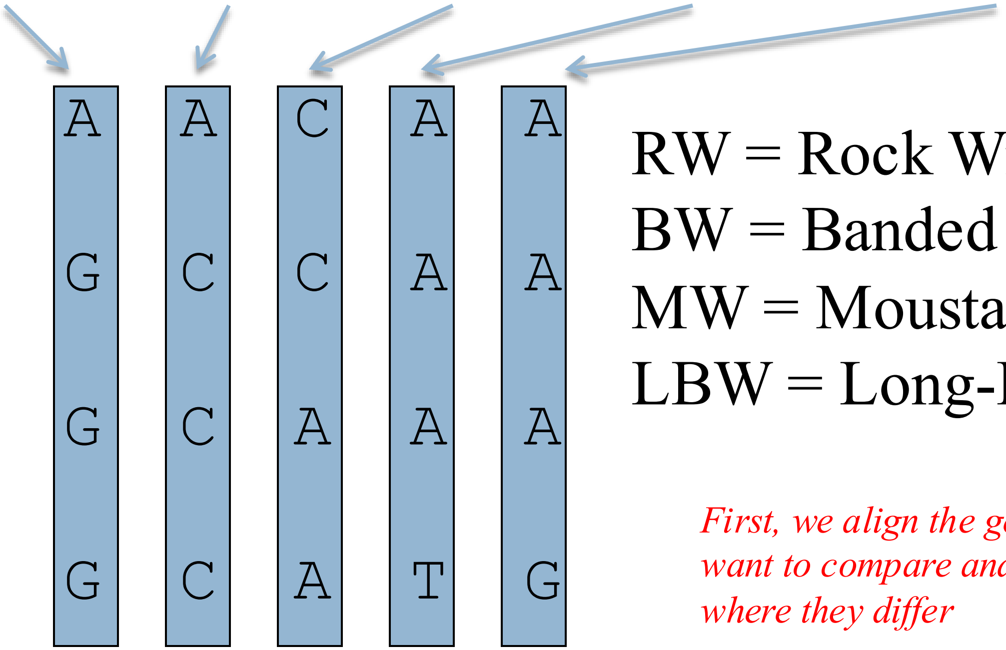
- Many methods, some quite sophisticated
- Simplest is to apply the principle of parsimony:
  - Occam's Razor - All other things being equal, the simplest solution is the best.
- Can be done on Morphological or Molecular data

THE PARSIMONY CRITERION

Select the tree that implies the minimum number of character changes summed across all characters of interest

Let’s use molecular data

RW	A A C G A T T C T A A A G G A T T
BW	A G C G C T T C T A A A G G A T T
MW	A G C G C T T A T A A A G G A T T
LBW	A G C G C T T A T A T A G G G T T

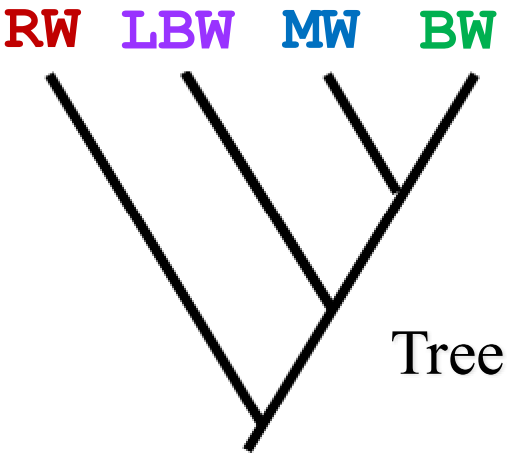


RW	A	A	C	A	A
BW	G	C	C	A	A
MW	G	C	A	A	A
LBW	G	C	A	T	G

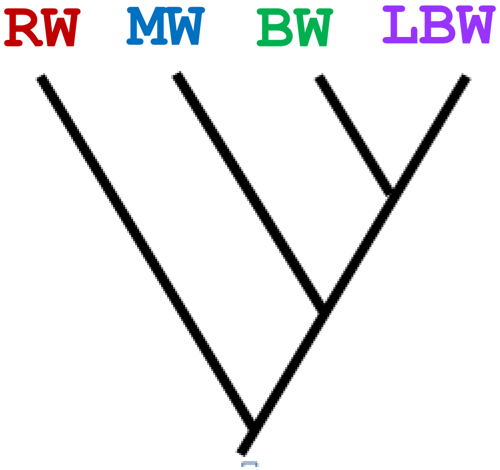
RW = Rock Wren (outgroup)
BW = Banded Wren
MW = Moustached Wren
LBW = Long-Billed Wren

First, we align the gene sequences that we want to compare and look for places where they differ

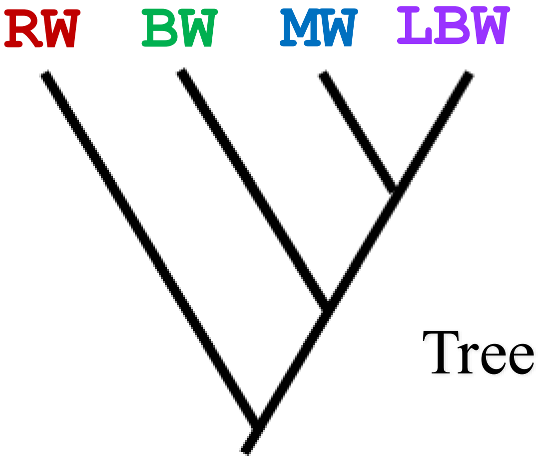
Three possible trees



Tree 1



Tree 2



Tree 3

We already know that RW is the outgroup.

There are three phylogenies (trees) that we would like to test.

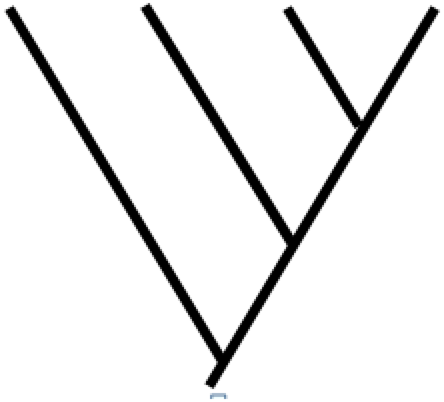
Only one of these can be correct.

Three possible trees

(RW,(LBW,(MW ,BW)));

Tree 1

(RW,(MW ,(BW , LBW)));



Tree 2

(RW,(BW, (MW , LBW)));

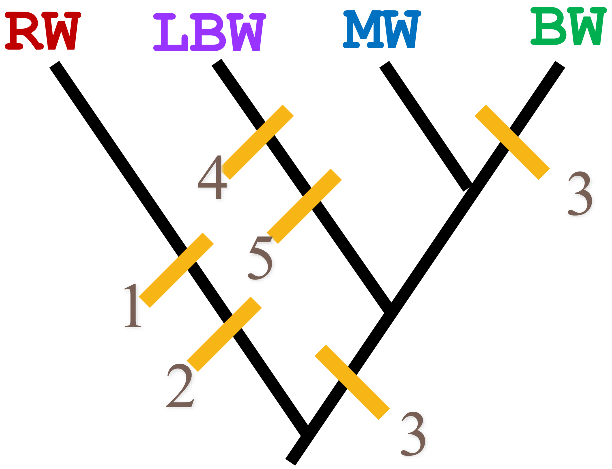
Tree 3

Here we've included the parentheses and commas we would have used to develop the top two trees.

Map the characters onto tree 1

These sequences from four different wren species are aligned. Each column (1, 2, 3…) can be considered a character trait. We will therefore analyze each sequence position ("character") separately.

	1	2	3	4	5
RW	A	A	C	A	A
BW	G	C	C	A	A
MW	G	C	A	A	A
LBW	G	C	A	T	G



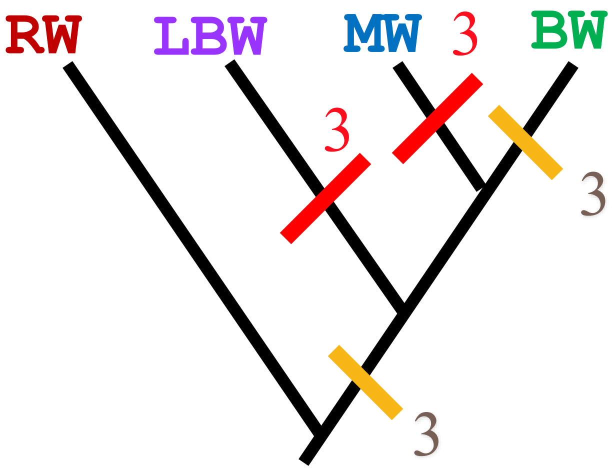
Let's map the remaining character changes

The number mapped above on the tree depicts the "character" change to the left. For 1, the rock wren "character" changed from a G (guanine) to an A (adenine). For 2, the rock wren "character" changed from a C (cytosine) to an A (adenine).

Total cost (length) = 6 steps

Actually there is more than one way to map character 3

	1	2	3	4	5
RW	A	A	C	A	A
BW	G	C	C	A	A
MW	G	C	A	A	A
LBW	G	C	A	T	G

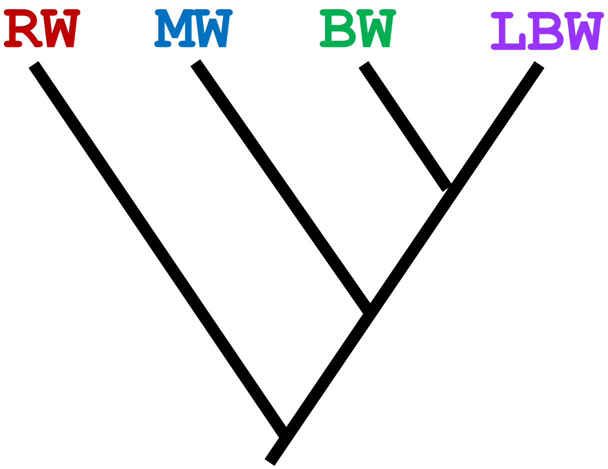


Either way the character contributes 2 steps to the overall cost

i.e. – you can map this character change in different ways, each for the same cost (two separate changes needed).

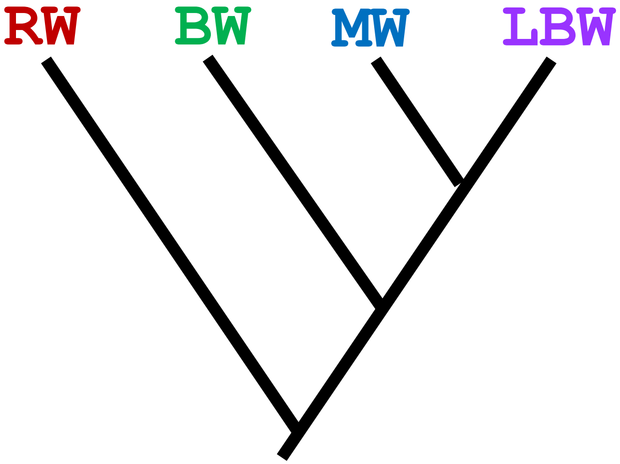
Map the characters onto tree 2

	1	2	3	4	5
RW	A	A	C	A	A
BW	G	C	C	A	A
MW	G	C	A	A	A
LBW	G	C	A	T	G



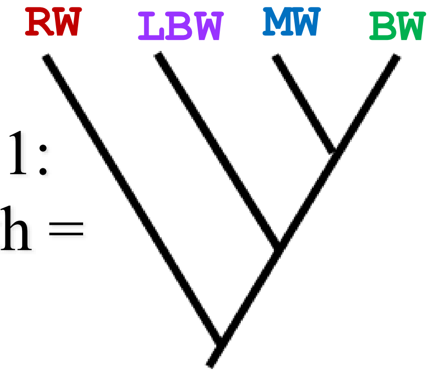
Map the characters onto tree 3

	1	2	3	4	5
RW	A	A	C	A	A
BW	G	C	C	A	A
MW	G	C	A	A	A
LBW	G	C	A	T	G

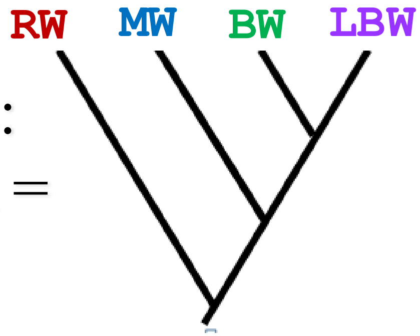


What was the cost of each tree?

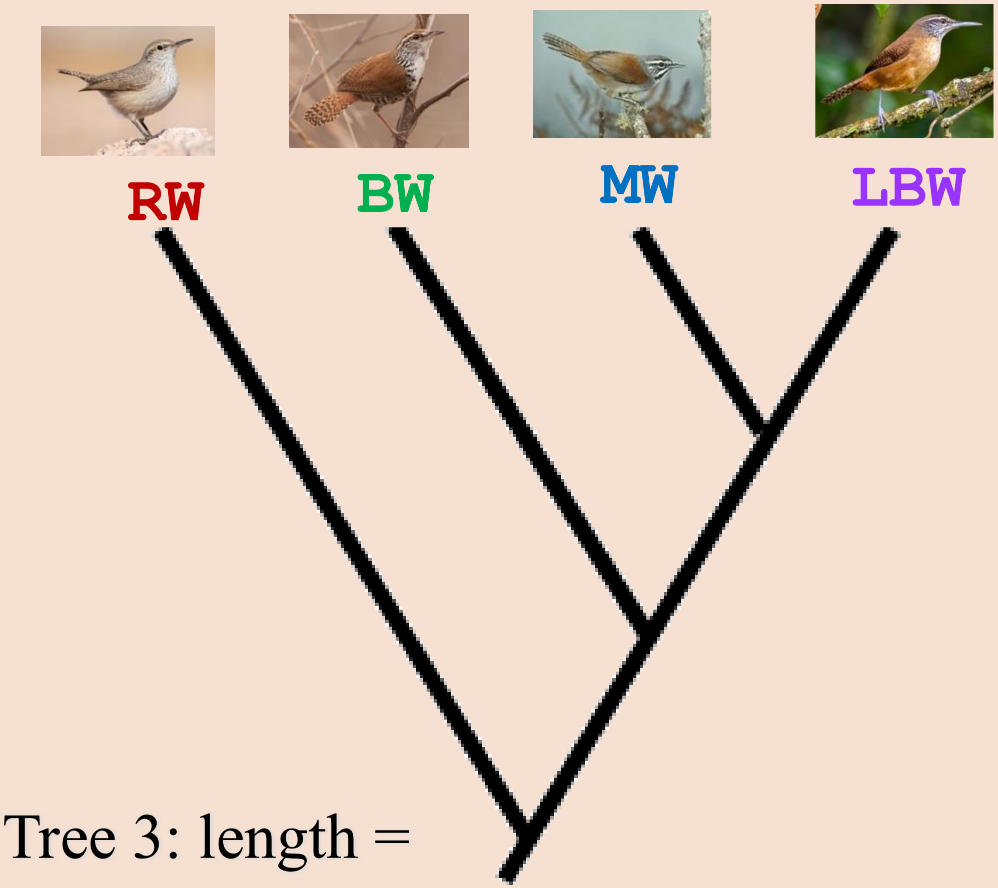
Tree 1:
length =



Tree 2:
length =



Tree 3 has the lowest cost, is therefore the most parsimonious tree and the one that we believe to be the most likely tree based on Occam's Razor.



Tree 3: length =

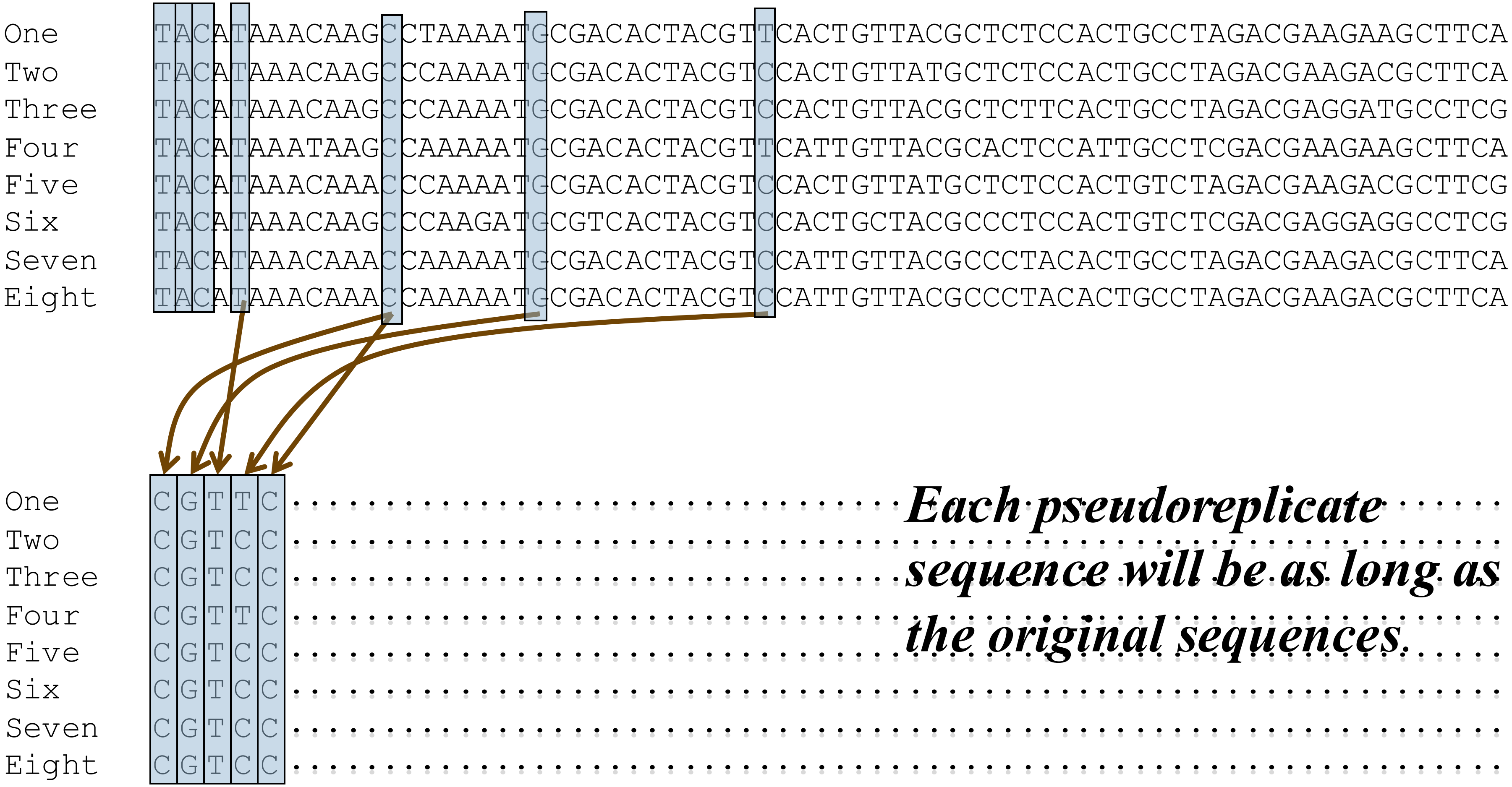
How do we evaluate uncertainty?

- One commonly used method: bootstrap analysis
- Bootstrapping is a resampling technique that allows us to make inferences about a dataset that may be, statistically, too small to do so otherwise.
- Bootstrapping is a way of testing the reliability of the dataset. It creates a pseudoreplicate dataset by resampling.
- We are bootstrapping our gene sequences to expand our dataset by making what are called pseudo-replicates.

How do we evaluate uncertainty?

- Repeated random sampling of *n* objects (in our case A/T/C/G) from the entire data set may result in a selection of groups where some objects can be incorporated into the same random sample several times while others may never be chosen.

Bootstrapping
Create "pseudoreplicate" data sets



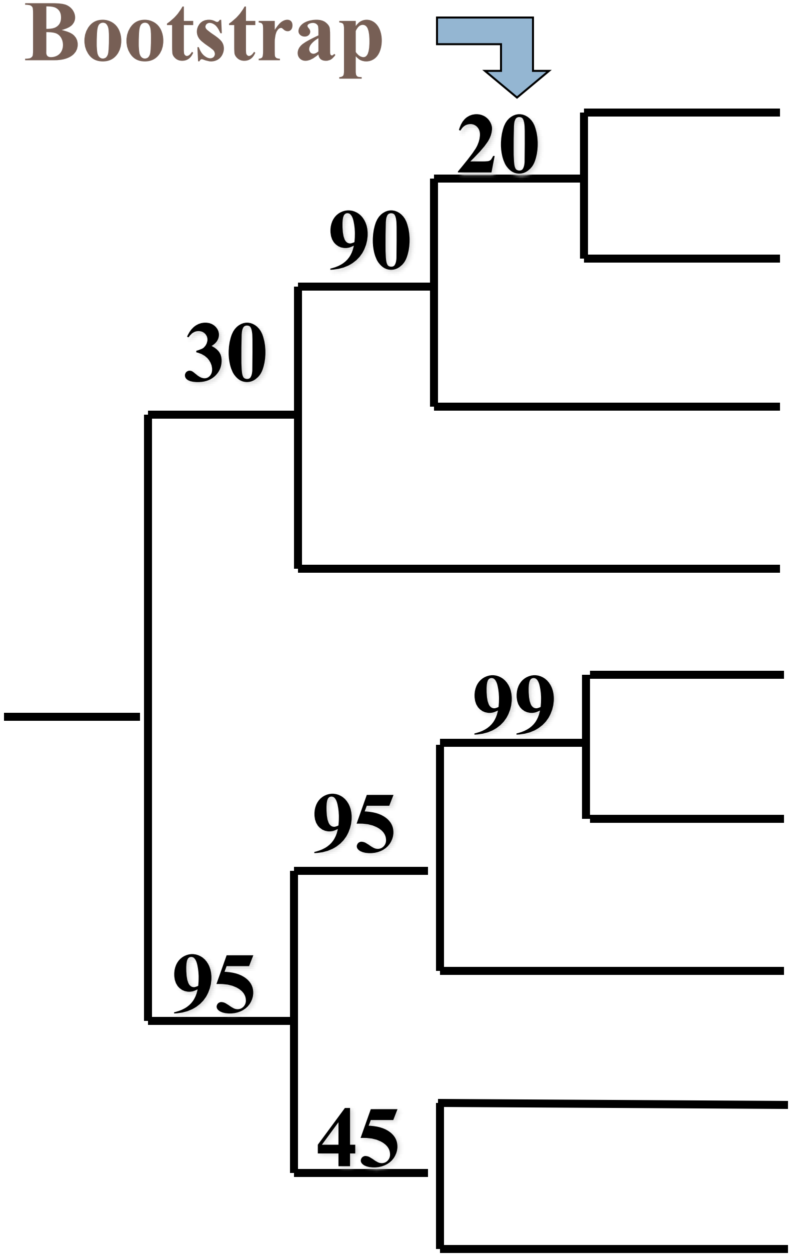
```
One      TACATAAACAAGCCTAAAATGCGACACTACGTTCACTGTTACGCTCTCCACTGCCTAGACGAAGAAGCTTCA
Two      TACATAAACAAGCCCAAAATGCGACACTACGTCCACTGTTATGCTCTCCACTGCCTAGACGAAGACGCTTCA
Three    TACATAAACAAGCCCAAAATGCGACACTACGTCCACTGTTACGCTCTTCACTGCCTAGACGAGGATGCCTCG
Four     TACATAAATAAGCCAAAAATGCGACACTACGTTCATTGTTACGCACTCCATTGCCTCGACGAAGAAGCTTCA
Five     TACATAAACAAACCCAAAATGCGACACTACGTCCACTGTTATGCTCTCCACTGTCTAGACGAAGACGCTTCG
Six      TACATAAACAAGCCCAAGATGCGTCACTACGTCCACTGCTACGCCCTCCACTGTCTCGACGAGGAGGCCTCG
Seven    TACATAAACAAACCAAAAATGCGACACTACGTCCATTGTTACGCCCTACACTGCCTAGACGAAGACGCTTCA
Eight    TACATAAACAAACCAAAAATGCGACACTACGTCCATTGTTACGCCCTACACTGCCTAGACGAAGACGCTTCA

One      CGTTC..........................................................
Two      CGTCC..........................................................
Three    CGTCC..........................................................
Four     CGTTC..........................................................
Five     CGTCC..........................................................
Six      CGTCC..........................................................
Seven    CGTCC..........................................................
Eight    CGTCC..........................................................
```

Each pseudoreplicate sequence will be as long as the original sequences.

Bootstrap procedure

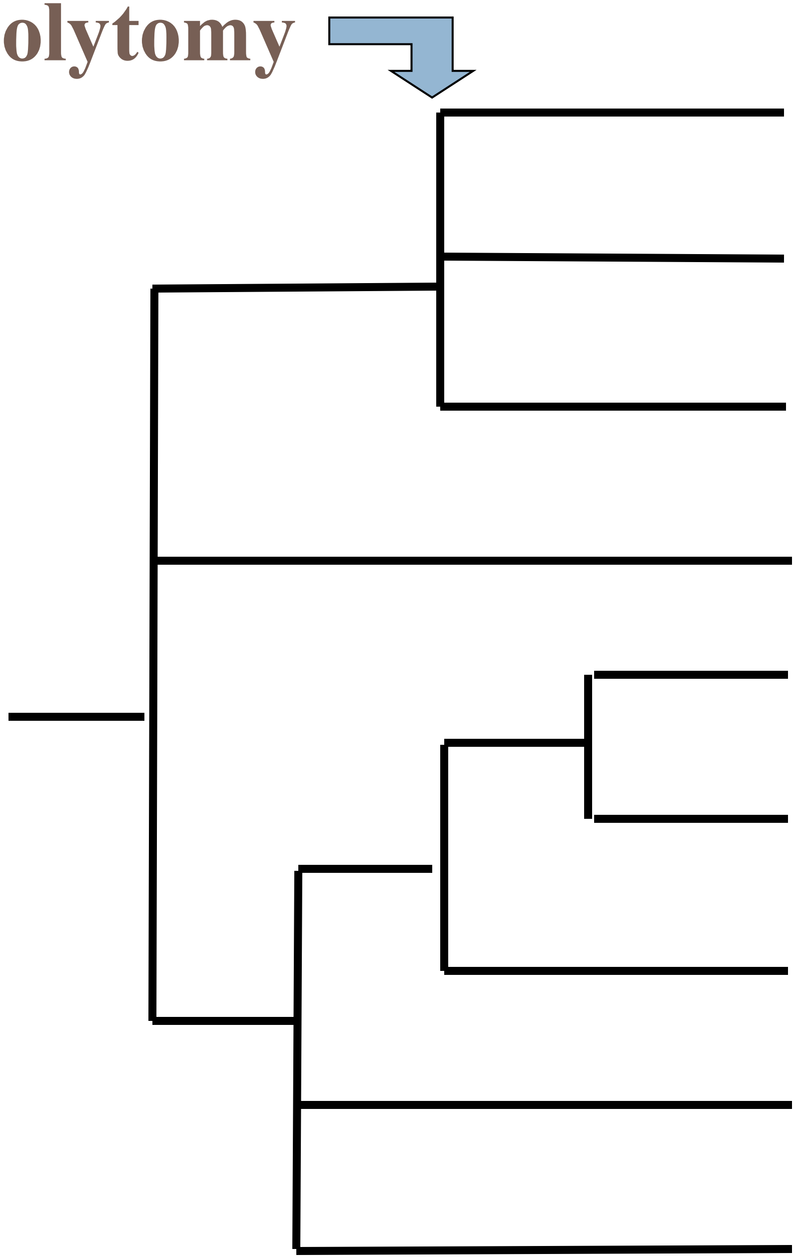
- Make many pseudo-replicates
- Find the most parsimonious (MP) trees for each
- Calculate the proportion of the time that the MP trees contain a clade (“bootstrap percentage”)
- Clades with high bootstrap support are better supported

Bootstrap



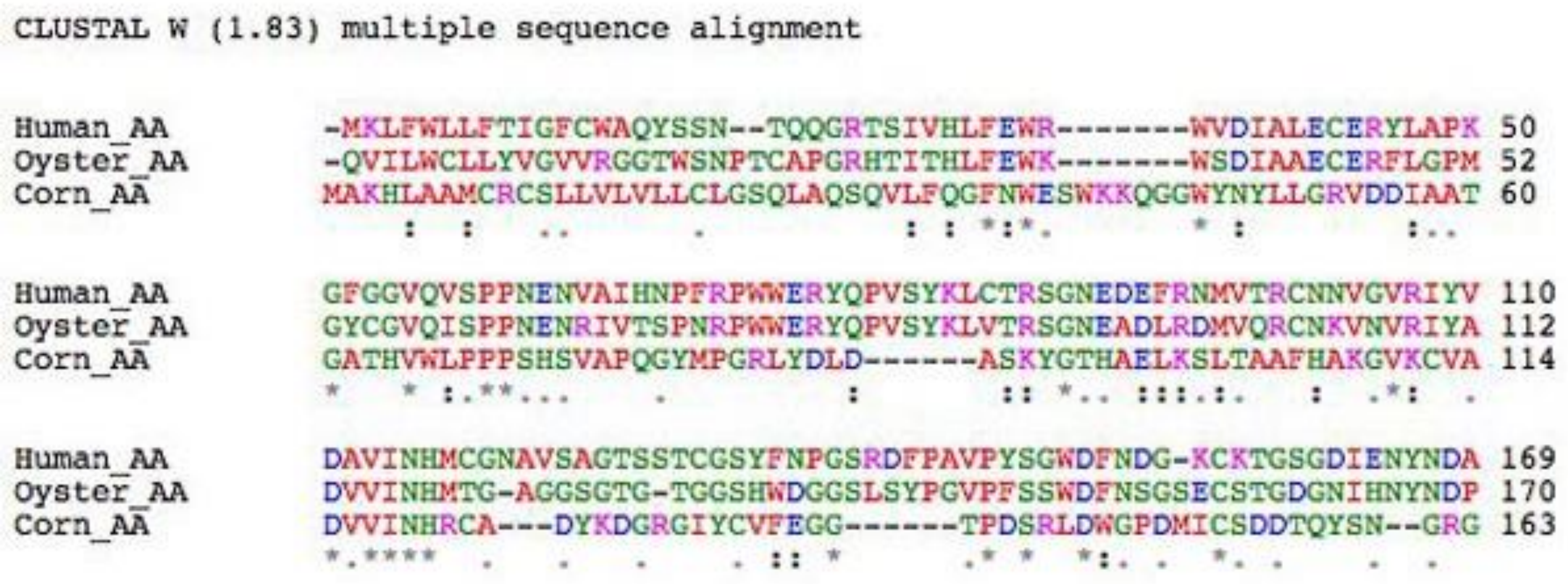
What we get

Polytomy



What we hypothesize

Amino acid alignment

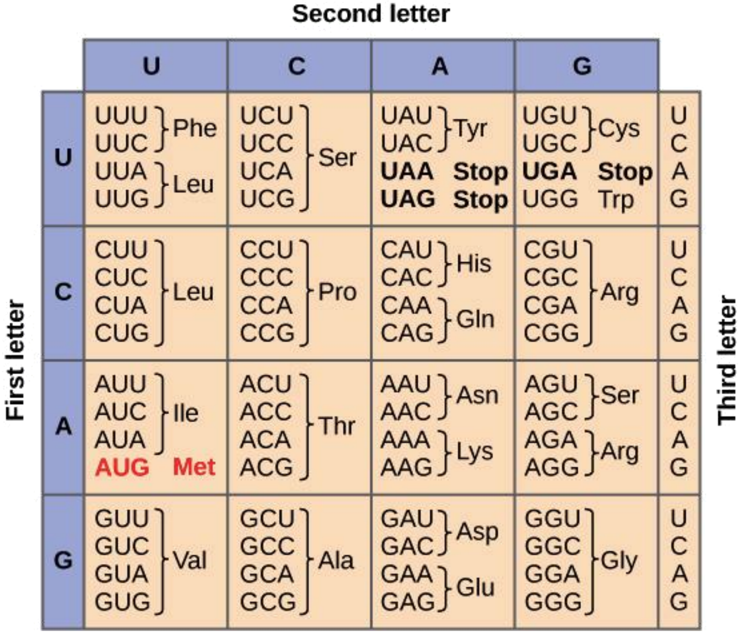


```
CLUSTAL W (1.83) multiple sequence alignment


Human_AA        -MKLFWLLFTIGFCWAQYSSN--TQQGRTSIVHLFEWR-------WVDIALECERYLAPK 50
Oyster_AA       -QVILWCLLYVGVVRGGTWSNPTCAPGRHTITHLFEWK-------WSDIAAECERFLGPM 52
Corn_AA         MAKHLAAMCRCSLLVLVLLCLGSQLAQSQVLFQGFNWESWKKQGGWYNYLLGRVDDIAAT 60
                    :  :   ..      .          : : *:*.       * :        :..

Human_AA        GFGGVQVSPPNENVAIHNPFRPWWERYQPVSYKLCTRSGNEDEFRNMVTRCNNVGVRIYV 110
Oyster_AA       GYCGVQISPPNENRIVTSPNRPWWERYQPVSYKLVTRSGNEADLRDMVQRCNKVNVRIYA 112
Corn_AA         GATHVWLPPPSHSVAPQGYMPGRLYDLD------ASKYGTHAELKSLTAAFHAKGVKCVA 114
                *   * :.**...    .         :       :: *.. :::.:.   :  .*:  .

Human_AA        DAVINHMCGNAVSAGTSSTCGSYFNPGSRDFPAVPYSGWDFNDG-KCKTGSGDIENYNDA 169
Oyster_AA       DVVINHMTG-AGGSGTG-TGGSHWDGGSLSYPGVPFSSWDFNSGSECSTGDGNIHNYNDP 170
Corn_AA         DVVINHRCA---DYKDGRGIYCVFEGG------TPDSRLDWGPDMICSDDTQYSN--GRG 163
                *.****  .   .   .    . :: *      .* *  *:. .  *. .    .  .
```

Condons

- Cells decode mRNAs by reading their nucleotides in groups of three, called **codons**. Here are some features of codons:
  - Most codons specify an amino acid
  - Three "stop" codons mark the end of a protein
  - One "start" codon, AUG, marks the beginning of a protein and also encodes the amino acid methionine

The genetic code table

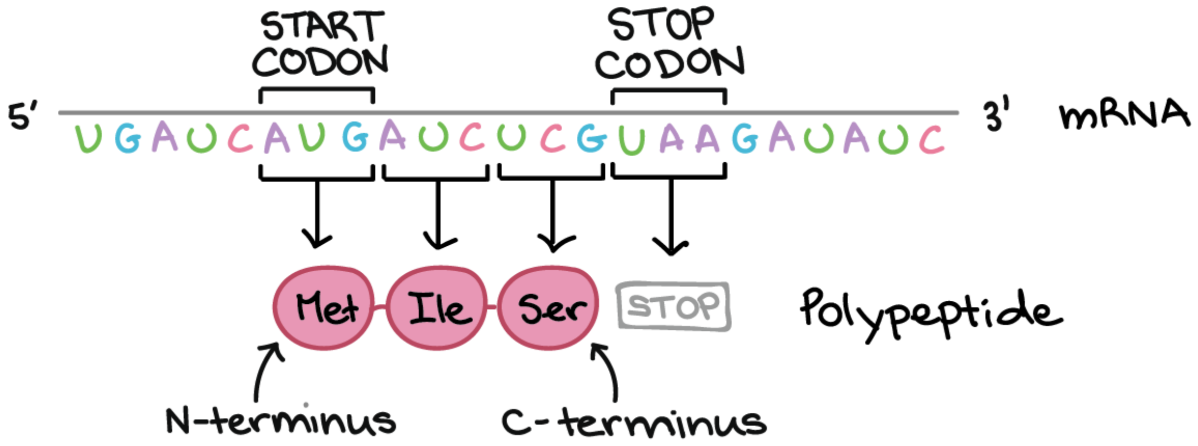


Second letter

First letter	U	C	A	G	Third letter
U	UUU UUC Phe UUA UUG Leu	UUU UCC UCA UCG Ser	UAU UAC Tyr **UAA Stop** **UAG Stop**	UGU UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU CAC His CAA CAG Gln	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile **AUG Met**	ACU ACC ACA ACG Thr	AAU AAC Asn AAA AAG Lys	AGU AGC Ser AGA AGG Arg	U C A G
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU GAC Asp GAA GAG Glu	GGU GGC GGA GGG Gly	U C A G

Condons

- Codons in an mRNA are read during translation, beginning with a start codon and continuing until a stop codon is reached.



MEGA Analysis

- Conduct an analysis in MEGA on *Cytochrome c oxidase I* molecular sequences