

Section 1: Phylogenetic tree basics-gathering genomic data and creating a tree

Why build a tree? Building a phylogenetic tree allows us to hypothesize evolutionary relationships between organisms. Unexpected relationships may be identified and expected relationships may be supported.

Phylogenetic trees are created using a number of different methods but the basic vocabulary is the same. Figure 1 to the right shows the fundamental structure and nomenclature of phylogenetic trees. Nodes are located at the beginning and end of each branch and represent taxonomic units. Terminal nodes represent existing species while internal nodes represent ancestors. The branch defines the lineages between the taxa at the nodes and represents gradual sequence changes from now extinct organisms. If scaled, the branch length represents the number of changes that have occurred in the branch and infers evolutionary time.

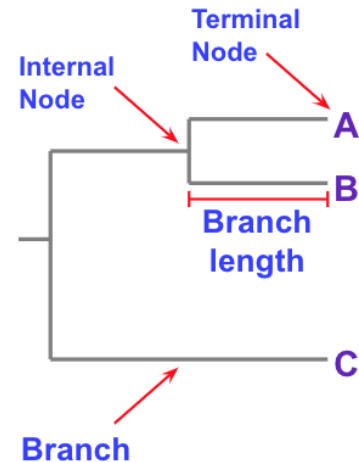


Figure 1. Phylogenetic Tree

Phylogenetic trees can be rooted or unrooted (Figure 2). Rooted trees have a designated root node, which is a known common ancestor or a distant outgroup. Rooted trees indicate the relationships between taxa and the direction of evolutionary change. If a common ancestor is not known then an unrooted tree is produced. An unrooted tree specifies relationships among the taxa but cannot say anything about the series of evolutionary events that led to existing species.

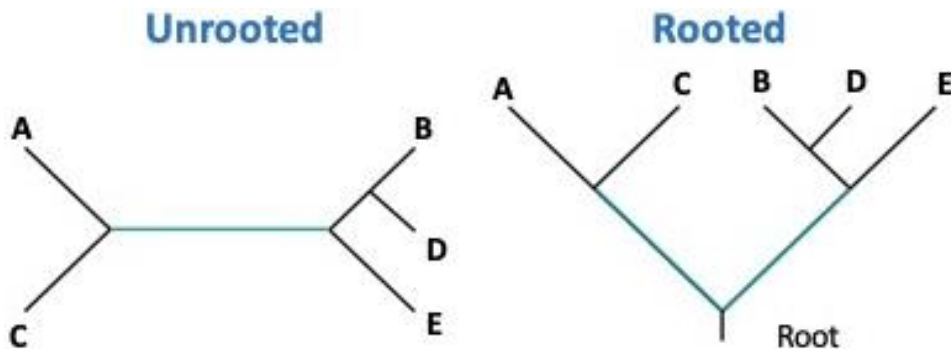


Figure 2. Representative trees that are either unrooted (left) or rooted by a common ancestor (right)

Question Set 1:

1.1 In the unrooted tree in Figure 2, which organisms are most distantly related? Which are most closely related?

1.2 In the rooted tree in Figure 2, which organisms are most distantly related? Which are most closely related?

1.3 For bacteriophage analysis, would you use an unrooted or rooted tree? Why?

Nucleic acid and protein sequence data can be compared directly by aligning the sequences and identifying differences and similarities, but how does this relate to evolutionary

relationships? Changes in the sequence of nucleotides or the amino acids occurs over time so the number and types of changes can be used to measure evolutionary time and predict relationships between organisms.

Phylogenetic trees are computed from molecular sequence data either directly or indirectly. Parsimony is an example of a direct method where a tree is produced that requires the fewest number of evolutionary changes. Indirect analysis involves first computing a distance matrix and then applying a method such as neighbor-joining to create a tree based on finding pairs of sequences (neighbors) that have the shortest branch distances. Recall that branch distance relates to evolutionary time!

Sequence data can be converted to distance data by measuring how many changes occurred between two sequences. Then the distance data is used to produce a distance matrix and then through neighbor-joining to produce a tree. Neighbor-joining is useful for rapidly analyzing large data sets that have low levels of sequence divergence.

Exercise 1: Construct your first tree using data in a distance matrix

Using the data in the distance matrix to the right, draw an unrooted tree showing the relationships between species A, B, C, D, and E.

To get started, find the two closest species and then join them. Work out how far their common ancestor is from the nearest species. Add the third species to the tree and then repeat for the fourth and fifth species.

Distance Matrix

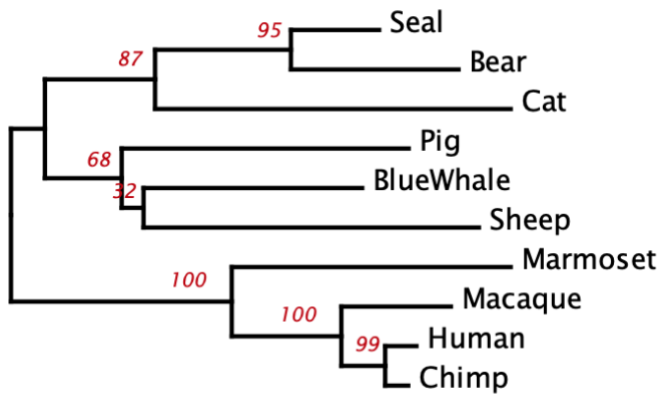
	A	B	C	D	E
A	0				
B	3	0			
C	5	6	0		
D	5	6	6	0	
E	7	8	8	4	0

Question Set 2:

1.4 Take a screen shot of your tree and paste in your worksheet.

We can check how robust a phylogenetic tree is by comparing trees produced using different methods and by using bootstrapping, which is a measure of data consistency.

Essentially to bootstrap a tree, the phylogeny program resamples the data (1000 times), ending up with variable trees, and variations in branches are noted as percentages. The percentages on the branches in red in the figure below indicate percent of time that branch occurred in a tree when the data is resampled. 100% (or 1.0 for some programs) indicates that the branch was always present in every tree!



In this example, the branch connecting bear and seal has a bootstrap value of 95% so we can be confident that these two organisms are closely related.

Figure 3. Example rooted tree with bootstrap values for a gene found in animals

Question Set 3:

- 1.5 Which branches are you the most confident in for the figure above? Which are you least confident in?
- 1.6 Which branch are you the least confident in and what conclusion can you draw from this branch?
- 1.7 Is the tree above rooted or unrooted? How were you able to determine that it is rooted or unrooted?
- 1.8 Which two related organisms have the shortest branch length, and what does this indicate?