Name________________________

Introduction to Bioinformatics: Accessing the NCBI Databases

By Shawn Lester

(12 points)

The purpose of this exercise is to introduce to you and familiarize you with the National Center for Biotechnology Information (NCBI) web sites. These web sites contain nucleotide and protein databases containing a vast amount of sequencing data that is increasing every single day. As more and more genes, genomes, and proteins are being sequenced by scientists from all over the world, these databases represent a central repository for all of these data which can be accessed by anyone on the planet.

You will be one of these people and you will search the databases to find homologous amino acid and nucleotide sequences in order to identify an unknown bacterium. To start with, your instructor will provide you with an unknown sequence (either amino acids or nucleotides) so that you may become familiar with how to search the databases. Later, you may actually be isolating and sequencing genes from living organisms! (To be decided later)

The first thing to do is to go to the NCBI web site. You can either type the word blast into your browser's search bar or use this address: http://blast.ncbi.nlm.nih.gov . As you will see, there is a tremendous amount of information located here and it is not that easy to interpret. We are not trying to make you sequencing experts. We want you to become familiar with the research tools that are available and hopefully you will begin to see the potential of what can be learned from these genetic data. When we perform our searches for homologous sequences, we will use the default settings. If you were an expert, you could change a variety of criteria when performing your searches. You should also know that the algorithms used to align the various sequences are not the only ones but are the most commonly used algorithms. When you go to the web site, take some time to look around.

Part A&B = Part 1

Part A:

If you have to type in your sequence, don't worry. Typos are not that critical.

1. Go to the NCBI web site: http://blast.ncbi.nlm.nih.gov

2. We are going to start practicing with an amino acid sequence for an unknown protein. The letters used may not be familiar to you. See the reference at the end of this handout for an explanation of amino acid abbreviations.

```
MYYLKNTNFWMFGLFFFFYFFIMGAYFPFFPIWLHDINHISKSDTGIIFAAISLFSLLFQPLFGLLSDK
```

3. In the middle of the web page where it says "Basic BLAST Choose a BLAST program to run", click on the hyperlink "protein blast". This hyperlink takes you to the default protein BLAST search page. The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. (NCBI web site, 2008)

4. Scroll down near the bottom and check the box that says "Show results in a new window." This makes things easier when performing repeated searches.

Name________________________

5. On the BLAST page, type or 'copy and paste' your amino acid sequence into the box under "Enter Query Sequence". Scroll to the bottom and click on BLAST button. Your alignment results will appear in a new window. This may take several minutes or longer depending on the complexity of the sequence, user traffic, and the time of day.

6. As you scroll down the page, you will see a great deal of information. Answer the following questions: **(5 points)** (Some of the answers require you to look them up elsewhere on the internet.)
   a. How many 'Blast Hits' were found in response to your search query?________________

   b. What is the first organism listed that matches your search?________________________

   c. What enzyme was encoded by the amino acid sequence you searched?______________

   d. What is the function of this enzyme? (2 points)

   ____

7. Click on the "Descriptions" hyperlink. This expands the information for each listed organism produced by the search.

8. <u>**When**</u> you make a tree (in Part 2, to be done later), check the little box beside the name of 5 **different** organisms in the list. Example:

```
>☑  emb|CAA11068.1|  reverse transcriptase [Allium cepa]
```

   The organisms should at least be a different species. However, there are certain organisms for which there is very little diversity (HIV for example).

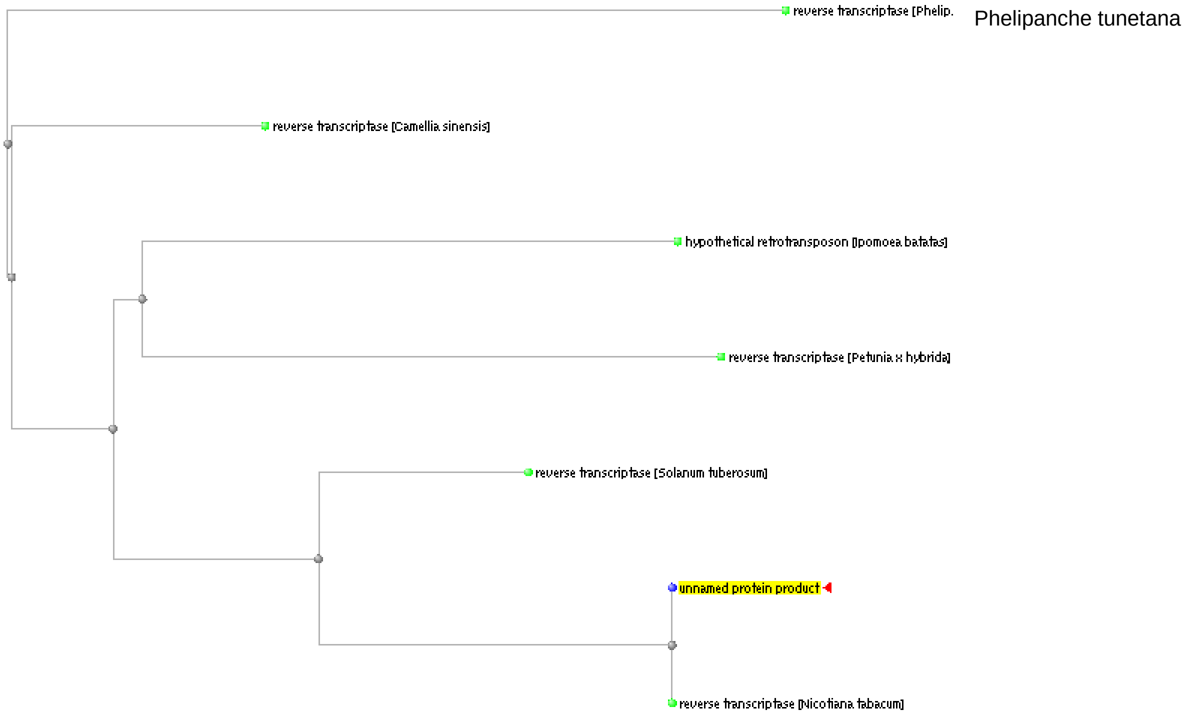
9. Scroll up a little and find the hyperlink <u>Distance tree results</u> and click on it. This will produce a phylogenetic tree showing distance (i.e. relatedness) of the five organisms you selected. **Do not** click on "<u>[Distance tree results]</u>" at the **top** of the page. This hyperlink automatically selects all of organisms produced from the original search which are displayed in a very complicated tree.

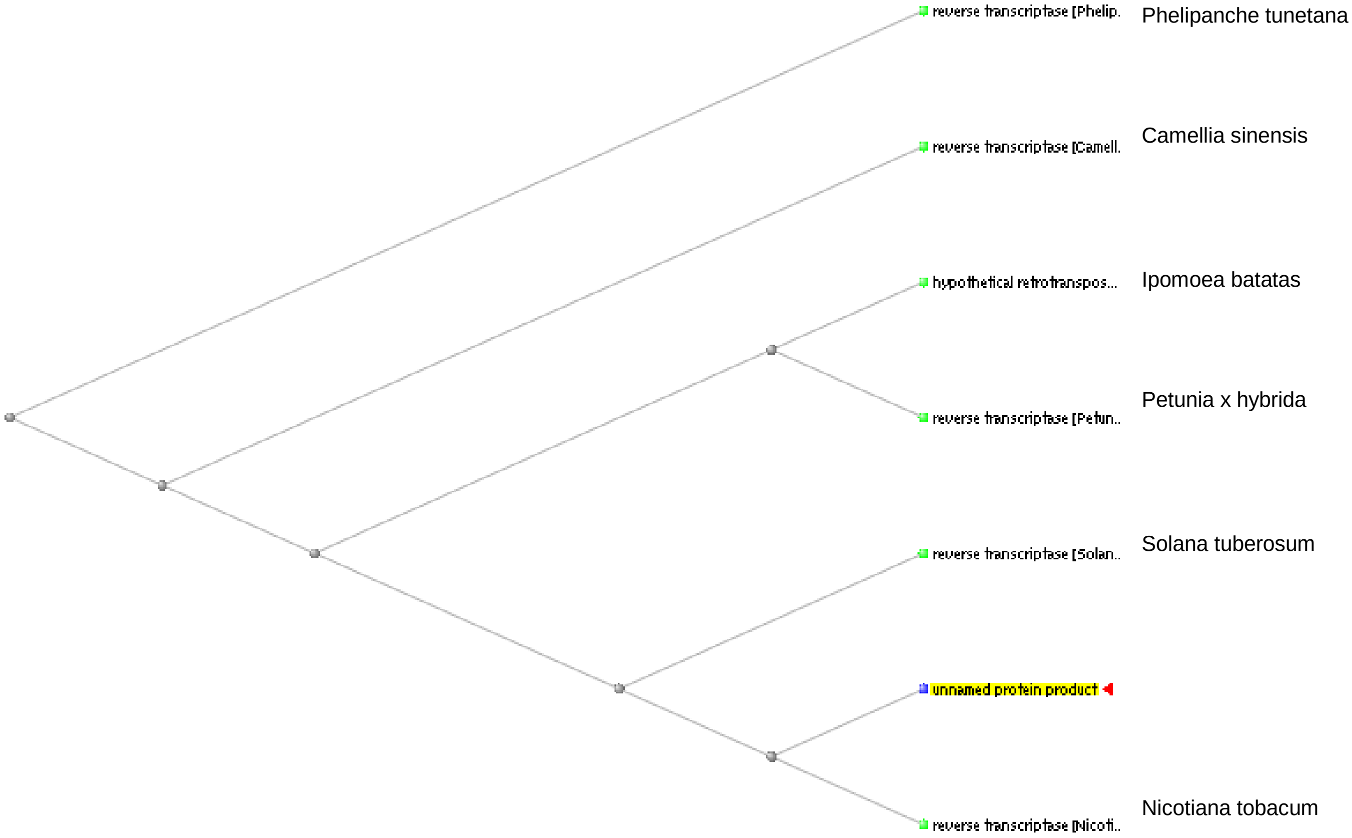
10. Above the tree, in the middle, is a box called "Collapse Mode". Select 'Show All'. This is very important. All of the names of the organisms may not be fully displayed unless you "Show All". When you make your tree, there are several ways to view it. The default view is "rectangle cladogram". Right click on the tree. Choose "Layout" to see the options. Sometimes "slanted cladogram" produces a tree that may be easier to interpret however the names may not be fully displayed. You should print both versions (rectangle and slanted) and <u>you may have to write the full names on the sheet</u>.

11. Right click on the phylogenetic tree, select "Save Picture As…" and save it to somewhere like the Desktop or select "Copy" and paste it into a Word document. Open the picture on your Desktop, then either print it or 'copy and paste' it to the document. (You may do this by copying and pasting the icon on the Desktop.)

Name___________________________

12. See example trees from a different search below. Note that the “unnamed protein product” highlighted in yellow **is** the organism on which you performed the blast search.


reverse transcriptase [Phelip.
Phelipanche tunetana
reverse transcriptase [Camellia sinensis]
hypothetical retrotransposon [Ipomoea batatas]
reverse transcriptase [Petunia x hybrida]
reverse transcriptase [Solanum tuberosum]
unnamed protein product
reverse transcriptase [Nicotiana tabacum]


reverse transcriptase [Phelip.
Phelipanche tunetana
reverse transcriptase [Camell.
Camellia sinensis
hypothetical retrotranspos...
Ipomoea batatas
reverse transcriptase [Petun..
Petunia x hybrida
reverse transcriptase [Solan..
Solana tuberosum
unnamed protein product
reverse transcriptase [Nicoti..
Nicotiana tobacum

Name________________________

13. Notice that the ‘unnamed protein product’ shown represents your original search protein. **Using the trees above,** which other organism’s reverse transcriptase is most closely related to *Nicotiana tobacum*? (The “unnamed protein product” is *Nicotiana tobacum.*) _____________________________**(1 point)**

Part B:

You have just completed your first BLAST search of a protein! Now we are going to do the exact same thing but with a nucleotide sequence.

1. Go back to the BLAST web page. (If you are starting over, repeat steps 1, 3, and 4 but at step 3, click on the word “nucleotide blast” instead of “protein blast”.) You can either click on the “blastn” tab at the top of the blast page or hit the back button on your web browser to go to the previous screen. Once there, click on “nucleotide blast”. Scroll down near the bottom and check the box that says “Show results in a new window.” For this search there are many possible databases to choose from. Find “Choose Search Set” and under “Database”, click the window box and choose: “Nucleotide collection (nr/nt)”. We want to use this general all purpose database and not search the human genome!

2. On the BLAST page, type or ‘copy and paste’ your nucleotide sequence into the box under “Enter Query Sequence”. Scroll to the bottom and click on BLAST button. Your alignment results will appear in a new window.

```
tgctcggggt ggaagggtcc taggccccga acaggggtca cgatgaggac cgtgcatagg cggattgacc caaaaaatgc
ccagccctct aagtagaggg cagaaaacct aagccgtctg gtggatggct cggctcgggg cgccgacgaa gggcgtggca
agctgcgata agccccggcg aggcgcaggc agccgtagaa ccggggattc ccgaatggga cttcctgcgg ctttgccgca
ctcccgtcag ggagggggaa cgcggggaat tgaaacatct tagt
```

3. As you scroll down the page, you will see a great deal of information. Answer the following questions: **(5 points)**

   a. How many ‘Blast Hits’ were found in response to your search query?_______________

   b. What organism is the best match based on the nucleotide sequence you searched?
   ____

   c. What gene or part of a gene was used to generate the search query?_______________

   d. Why is this particular gene commonly used for sequencing purposes? You must search the internet for this answer. (worth 2/5 points)

   ____

   ____

4. Click on the “Descriptions” hyperlink. This expands the information for each listed organism produced by the search.

5. **When** making a tree, check the little box beside the name of 5 different organisms in the list. Remember, where possible; choose 5 organisms that, at least, are a different species. Example:

Name________________________

```
>☑ gb|EU375368.1|  Corynebacterium diphtheriae isolate 4 16S ribosomal RNA gene,
partial sequence
```

6. Scroll up a little and find the hyperlink Distance tree results and click on it. This will produce a phylogenetic tree show distance (i.e. relatedness) of the five organisms you selected. Clicking on "[Distance tree results]" at the top of the page automatically selects all 100+ organisms which are displayed in a very complicated tree.

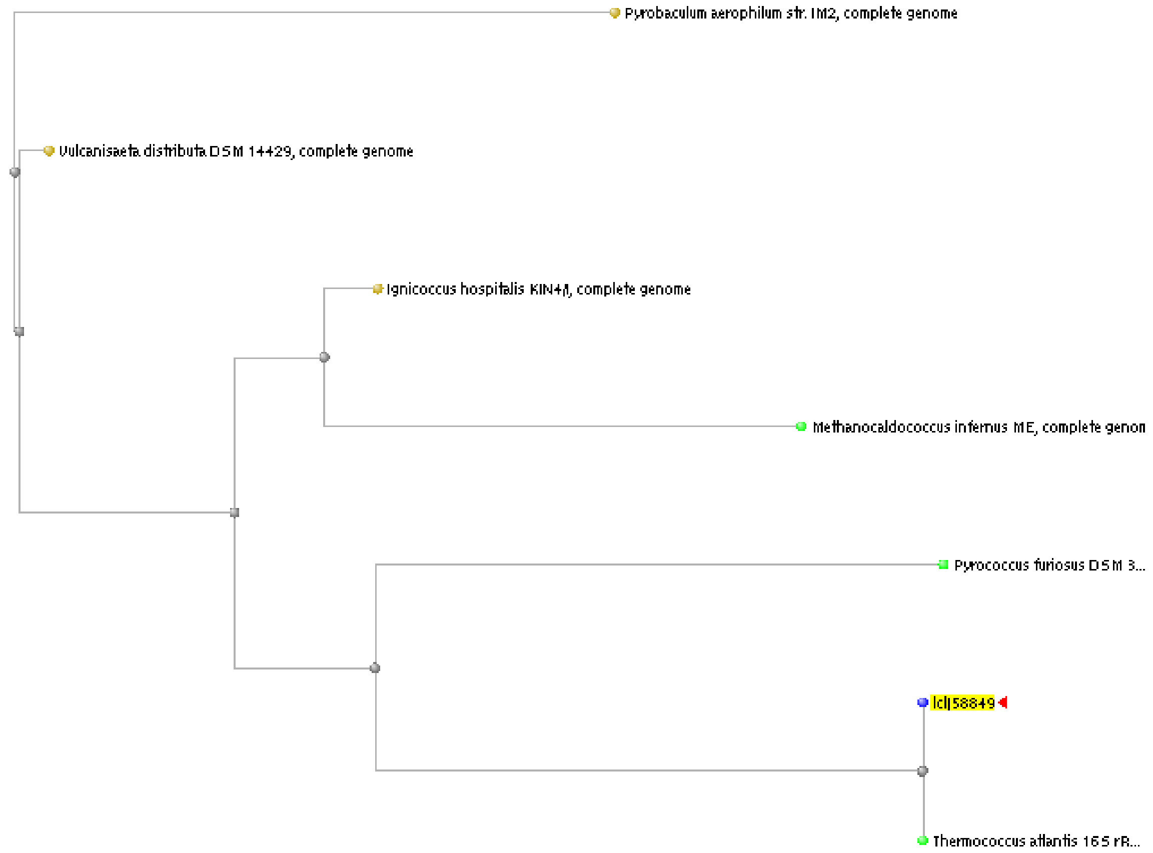
7. Above the tree, in the middle, is a box called "Collapse Mode". Select 'Show All'. This is very important. All of the names of the organisms may not be fully displayed unless you "Show All". When you make your tree, there are several ways to view it. The default view is "rectangle cladogram". Right click on the tree. Choose "Layout" to see the options. Sometimes "slanted cladogram" produces a tree that may be easier to interpret however the names may not be fully displayed. You should print both versions (rectangle and slanted) and you may have to write the full names on the sheet.

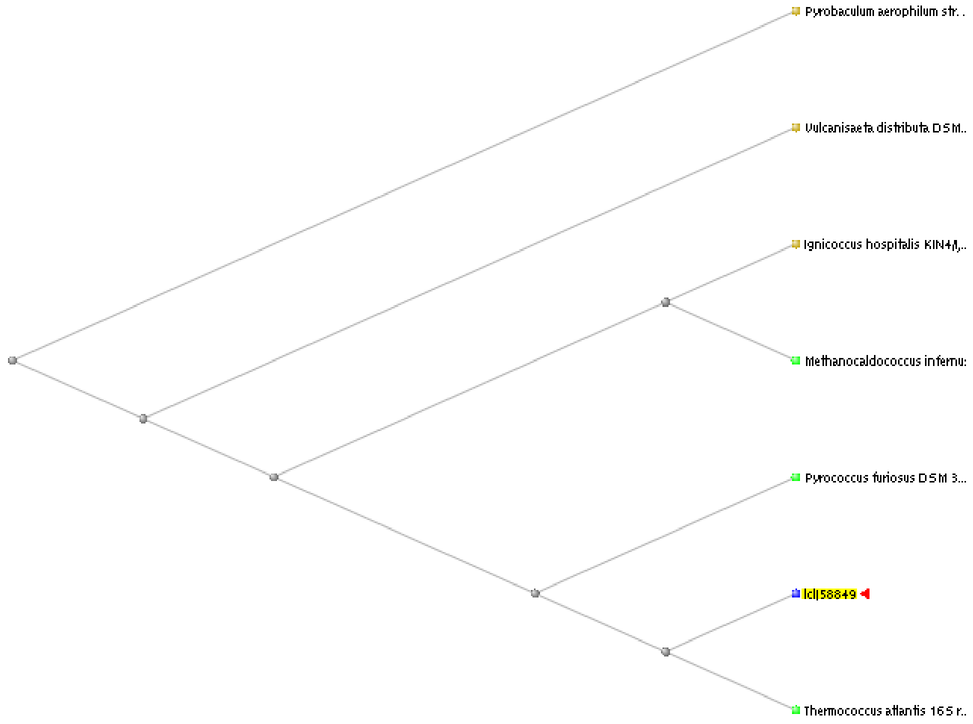
8. Right click on the phylogenetic tree, select "Save Picture As…" and save it to somewhere like the Desktop or select "Copy" and paste it into a Word document. Open the picture on your Desktop, then either print it or 'copy and paste' it to the document. (You may do this by copying and pasting the icon on the Desktop.)

9. See example trees on next page. Note that the "sequence" highlighted in yellow **is** the organism on which you performed the blast search.

10. Using the example trees on the next page, which organism is more closely related to *Thermococcus atlantis*?
____________________________**(1 point)**

Name_________________________


Pyrobaculum aerophilum str. IM2, complete genome
Vulcanisaeta distributa DSM 14429, complete genome
Ignicoccus hospitalis KIN4/I, complete genome
Methanocaldococcus infernus ME, complete genon
Pyrococcus furiosus DSM 3...
lcl|58849
Thermococcus atlantis 16S rR...


Pyrobaculum aerophilum str. .
Vulcanisaeta distributa DSM..
Ignicoccus hospitalis KIN4/I,..
Methanocaldococcus infernu:
Pyrococcus furiosus DSM 3...
lcl|58849
Thermococcus atlantis 16S r..

You now have the basic skill necessary to search for homologous protein or gene sequences. You can now apply these skills to help you identify your unknown bacteria. (To be decided later)

Name_______________________

References/Supplemental Material

Amino Acid Abbreviations:

A	alanine	P	proline
B	aspartate/asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	*U*	*selenocysteine*
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate/glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	*gap of indeterminate length*

Values:

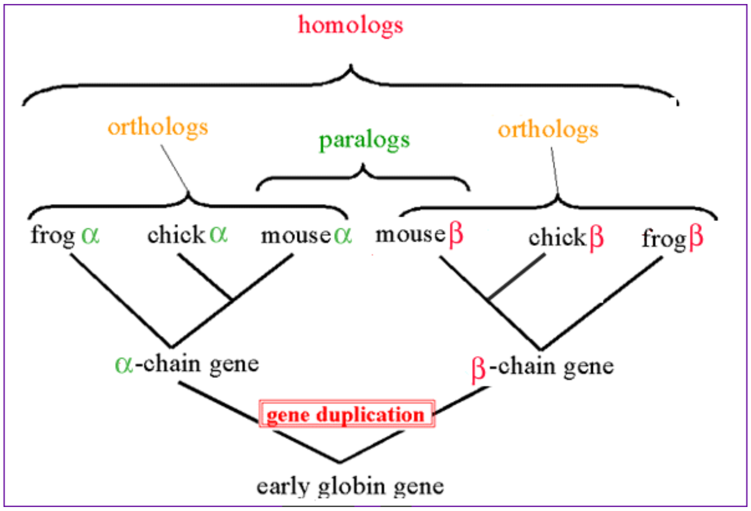
The E-value (expected) = a number that relates to an alignment match occurring by chance. Lower E-values represent better or more significant matches.

Terms:

Homology = Similarity attributed to descent from a common ancestor.

Orthology and Paralogy:

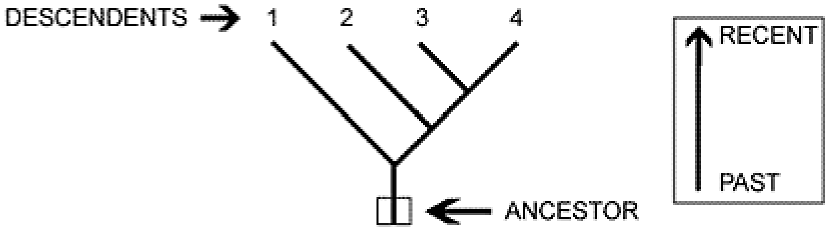
Homologous sequences. Orthologs and Paralogs are two types of homologous sequences. Orthology describes genes in different species that derive from a common ancestor. Orthologous genes may or may not have the same function. Paralogy describes homologous genes within a single species that diverged by gene duplication. (NCBI web site, 2008)



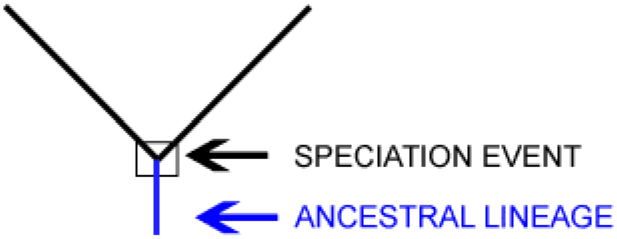
Name________________________

Understanding phylogenies (borrowed from evolution.berkeley.edu/evolibrary, 2008)

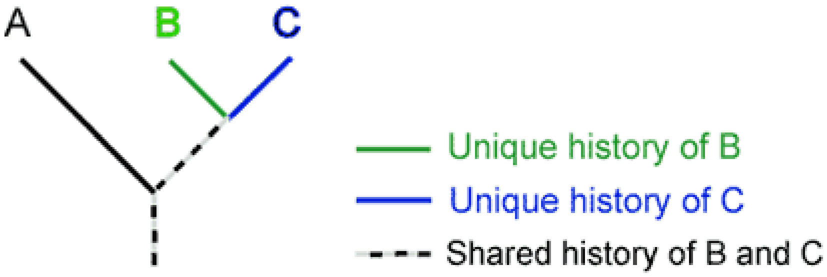
Understanding a phylogeny is a lot like reading a family tree. The root of the tree represents the ancestral lineage, and the tips of the branches represent the descendents of that ancestor. As you move from the root to the tips, you are moving forward in time.



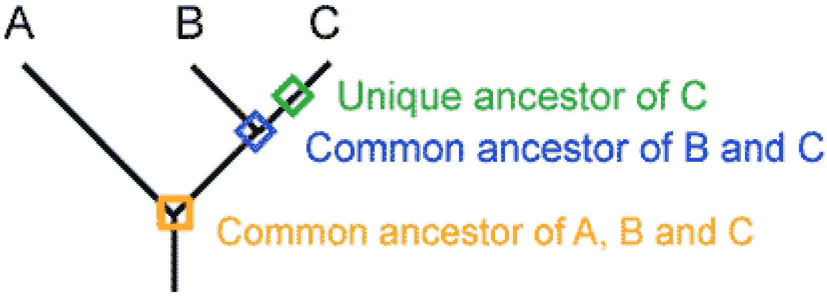
When a lineage splits (speciation), it is represented as branching on a phylogeny. When a speciation event occurs, a single ancestral lineage gives rise to two or more daughter lineages.



Phylogenies trace patterns of shared ancestry between lineages. Each lineage has a part of its history that is unique to it alone and parts that are shared with other lineages.

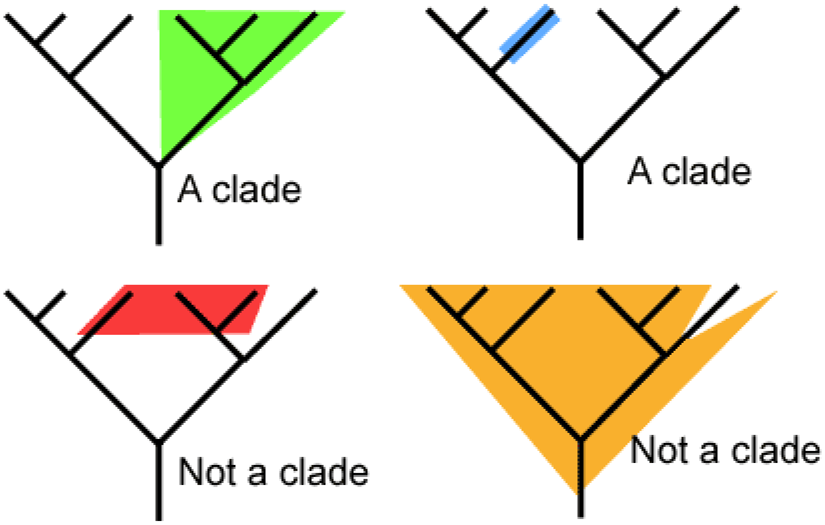


Similarly, each lineage has ancestors that are unique to that lineage and ancestors that are shared with other lineages — common ancestors.

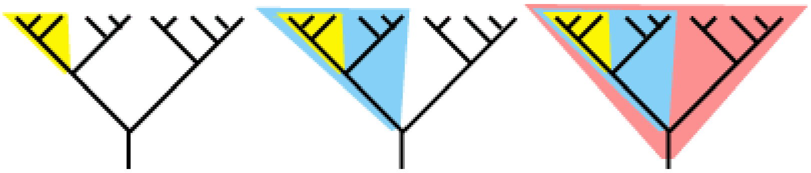


A clade is a grouping that includes a common ancestor and all the descendents (living and extinct) of that ancestor. Using a phylogeny, it is easy to tell if a group of lineages forms a clade. Imagine clipping a single branch off the phylogeny — all of the organisms on that pruned branch make up a clade.

Name________________________



Clades are nested within one another — they form a nested hierarchy. A clade may include many thousands of species or just a few. Some examples of clades at different levels are marked on the phylogenies below. Notice how clades are nested within larger clades.

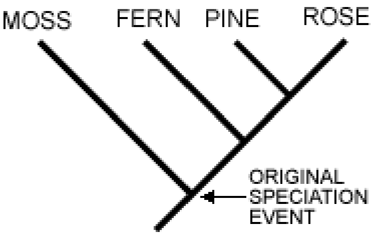


So far, we've said that the tips of a phylogeny represent descendent lineages. Depending on how many branches of the tree you are including however, the descendents at the tips might be different populations of a species, different species, or different clades, each composed of many species.

Trees, not ladders

Several times in the past, biologists have committed themselves to the erroneous idea that life can be organized on a ladder of lower to higher organisms. This idea lies at the heart of Aristotle's Great Chain of Being (see right).

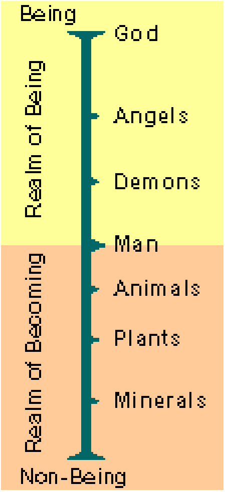
Similarly, it's easy to misinterpret phylogenies as implying that some organisms are more "advanced" than others; however, phylogenies don't imply this at all.



In this highly simplified phylogeny, a speciation event occurred resulting in two lineages. One led to the mosses of today; the other led to the fern, pine, and rose. Since that speciation event, both lineages have had an equal amount of time to evolve. So, although mosses branch off early on the tree of life and share many features with the ancestor of all land plants, living moss species are not ancestral to other land plants. Nor are they more primitive. Mosses are the cousins of other land plants.

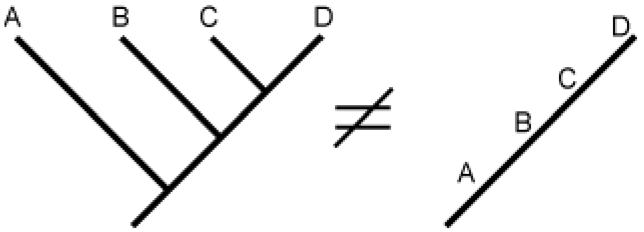
So when reading a phylogeny, it is important to keep three things in mind:

1. Evolution produces a pattern of relationships A B C D among lineages that is tree-like, not ladder-like.

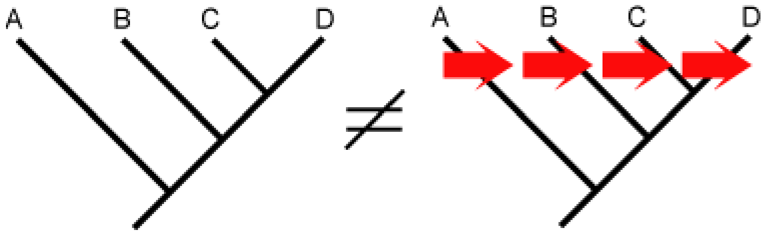


Aristotle's vision of a Great Chain of Being, above. We now know that this idea is incorrect.

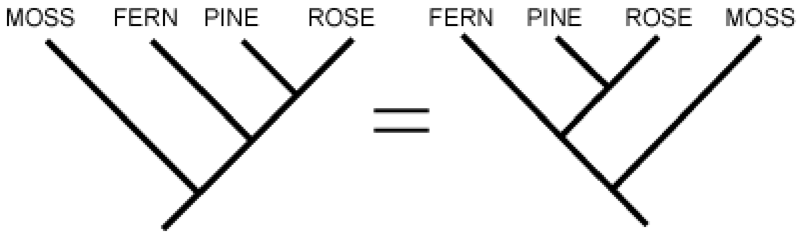
Name________________________



2. Just because we tend to read phylogenies from left to right, there is no correlation with level of "advancement."



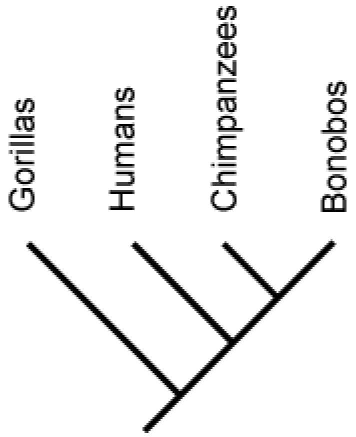
3. For any speciation event on a phylogeny, the choice of which lineage goes to the right and which goes to the left is arbitrary. The following phylogenies are equivalent:



Biologists often put the clade they are most interested in (whether that is bats, bedbugs, or bacteria) on the right side of the phylogeny.

Misconceptions about humans

The points described above cause the most problems when it comes to human evolution. The phylogeny of living species most closely related to us looks like this:



It is important to remember that:

1. Humans did not evolve from chimpanzees. Humans and chimpanzees are evolutionary cousins and share a recent common ancestor that was neither chimpanzee nor human.

2. Humans are not "higher" or "more evolved" than other living lineages. Since our lineages split, humans and chimpanzees have each evolved traits unique to their own lineages.