Using MEGA to Build Phylogenies

In this exercise, don't just check off the steps as you go! You need to focus on what each step is providing you – later you will need to use these resources for your own project. Take notes as needed, and respond to questions in italics on this sheet.

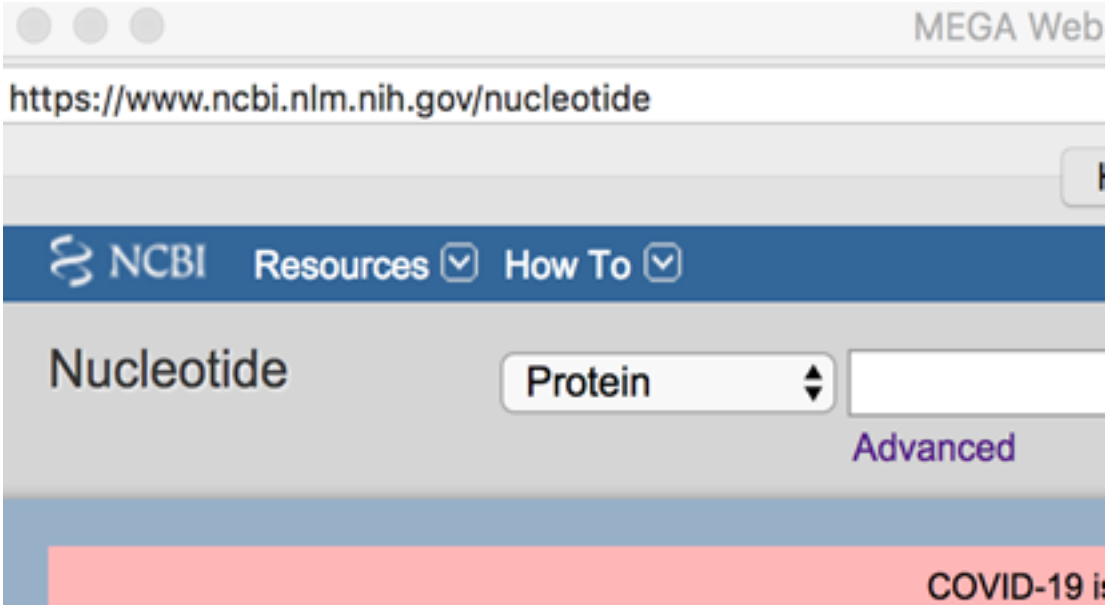
We will be using the gene "cytochrome b" to understand relationships between different eukaryotes. Look up "cytochrome b" – what is the gene used for? Why did we choose this gene to build a tree of eukaryotic life?

1. Download and install MEGA (http://www.megasoftware.net/) then open up the program.

2. Go to "Align" and click on "Query Databanks." This will take you to the NCBI website.

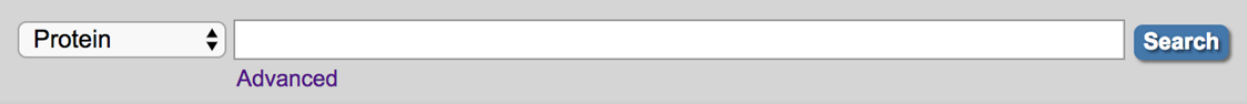
What is the NCBI and what does it do?

3. In the box to the left of the Search menu, switch from Nucleotide to Protein (you might have to change it in the URL box instead, plus be attentive because it seems to default to Nucleotide repeatedly). (Alternatively, click the blue search button after selecting protein. This seems to help it switch to focusing on proteins.)



Below the Search menu try selecting "advanced."

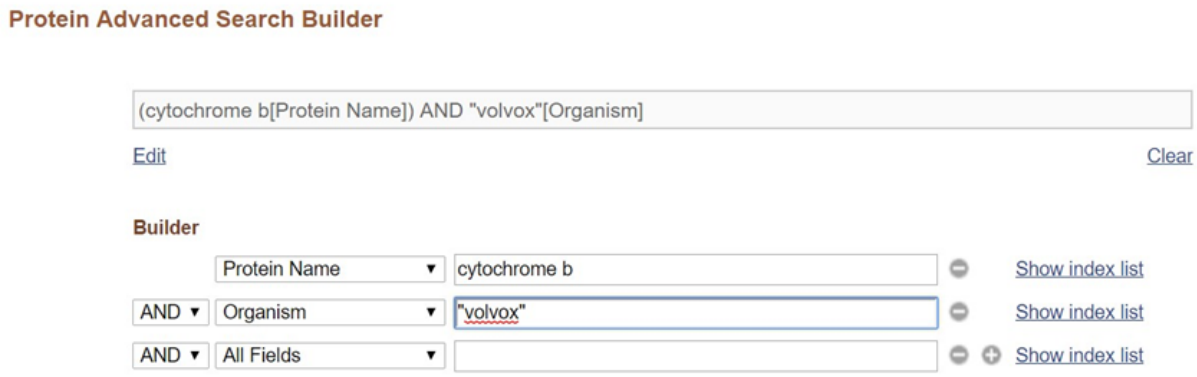
HINT: If it kicks you back into nucleotide search or does not do anything hit the blue "search" button while leaving the search box blank. This usually fixes the situation.



Then try entering Advanced again. You should see the “Protein Advanced Search Builder” come up.

Enter the following search criteria exactly as shown here:

Protein Name: cytochrome b
AND Organism: “volvox”

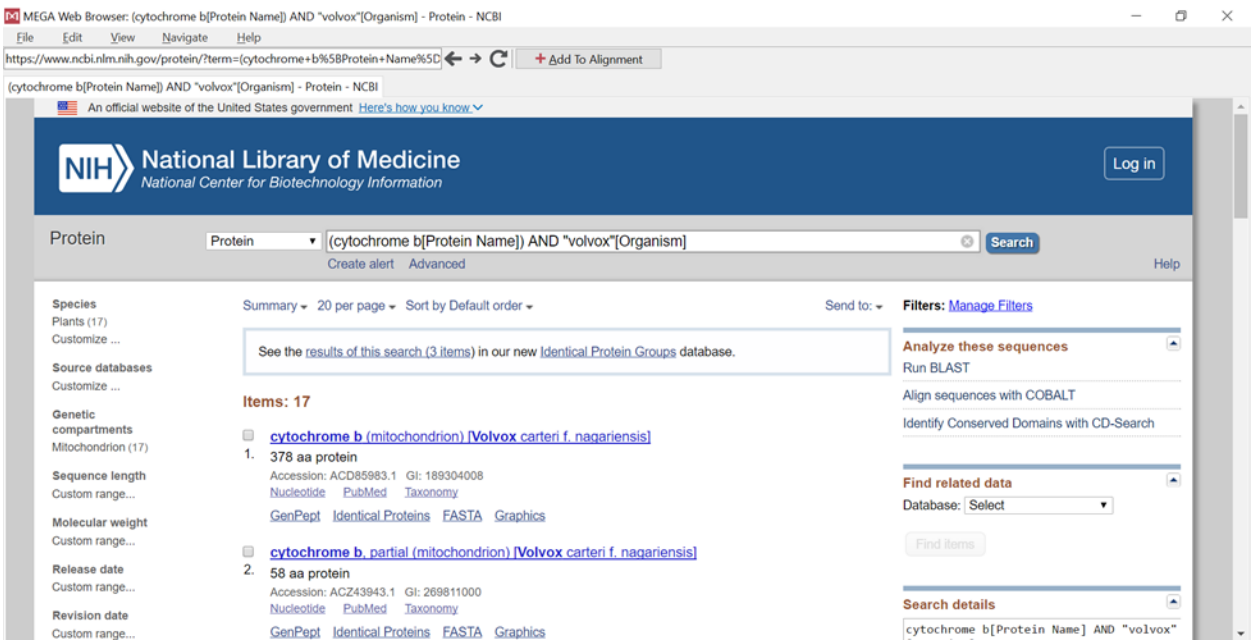


Click search.

HINT!!!! If at any point you get the message “well this is embarrassing” or another error, simply go back to your “Advanced Search” page where you should see your original search in the “History” chart there – DON’T redo your whole search!!!

Also if you get the “well this is embarrassing” error try hitting the back arrow a couple times.

4. As you look at your results, you should see multiple entries for “Volvox”, representing multiple different sequences or sequence fragments. The length of the sequence in the database is indicated below the protein name and is listed as a number of amino acids making up the protein. For example, the first one is 378 amino acids long. Click the check box next to the longest protein sequence you found. Make sure any species you add to your alignment has at least ~250 amino acids.



After you have selected the sequence, go to the "Summary" in the upper left and select the button for "FASTA (text)" – this will bring up a gray page with just the amino acid (protein) sequence. They may look something like MSGGKGGKA……..

Click "Add to Alignment" at the top of the page, which will add your sequences to the Alignment Explorer window in MEGA.

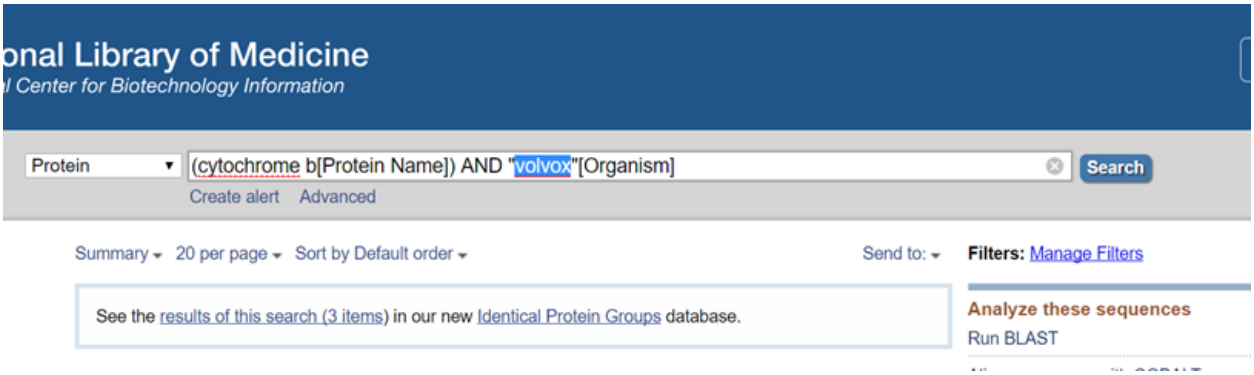
Once the Alignment window has opened, take this opportunity to edit the name so it doesn't include all the extraneous sequence information. Double click the name and then you can delete or re-write as desired. Indicate what clade the organism came from (initially, you'll be working with Protists, but we'll add Animals, Plants, and Fungi soon!)

For example, you might want to change the very long sequence title for Volvox to "Protist - Volvox".

You should see a page with lots of colors on it! Do not close this page since we will want to add more sequences from other groups before we are done. Once you have added your sequences to the alignment go back to the gray page and hit the back button. You should see your search still there.

5. Our goal today is to make a phylogenetic tree using the cytochrome b gene. To do that you will eventually need to locate 4-5 sequences each for animals, plants, fungi, and protists – for a total of 16-20 sequences. The sequence entries have the scientific name of the organism, not the common name.

To edit the search and add a new species to your alignment, highlight the name of the organism you previously searched for, replace it, and hit enter:



This database has the sequences for five of the six species of protists we observed in class today, so start by searching for the following searches and adding a sequence to the alignment:

1) Volvox:
(cytochrome b[Protein Name]) AND "volvox"[Organism]

2) Paramecium:
(cytochrome b[Protein Name]) AND "Paramecium"[Organism]

3) Amoeba:
(cytochrome b[Protein Name]) AND "Balamuthia"[Organism]

4) Euglena:
(cytochrome b[Protein Name]) AND "Euglena"[Organism]

Important Note: the Euglena and Spirostomum searches often come with only one sequence. Instead of popping up a list with check boxes, it takes you to a GenPept output. Once on that screen, click “GenPept” in the upper left hand corner, change it to “FASTA (text)” and then continue to add it to the alignment. You may have to click back a couple of times to get back to the search bar.

5) Spirostomum:
(cytochrome b[Protein Name]) AND "Spirostomidae"[Organism]

Finally, add an **outgroup** from the Domain Archaea. Archaea are the single-celled organisms that are most closely related to Eukaryotes and will be used as the “outgroup” to root the tree correctly.

6) Methanolobus:
(cytochrome b[Protein Name]) AND "Methanolobus"[Organism]

6. Now that you’ve added your five species of protists to your alignment, make a list of five diverse members of the other major clades of Eukaryotes: Animals, Plants, and Fungi. Search in Google and write down each organism’s binomial scientific name (so, Genus and species). For example, for Humans, you’d write “Homo sapiens”.

Animals are relatively easy to find in this database, but plants and fungi can be a bit more difficult sometimes. To help you out, we've prepared this helpful list of available plant and fungal species! Just make sure that you are selecting a wide variety of species in each taxon. You don't want to be selecting like 5 different willow trees (for example).

For each species, add the cytochrome b sequence to the alignment, repeating the above process to complete a new search each time and replacing the organism name with the next organism on your list. Sometimes you may find that the organism you'd like to add does not have sequence information. In that case, you'll have to come up with a replacement.

7. After all of your sequences are added to the Alignment Explorer (the page with all the bright colored amino acids lined up), you can edit the names of each sequence to remove the protein sequence information, leaving just the species name. Double click each species and edit the name to be more readable. For this project, it also may be helpful to add in the Kingdom for each organism. For example, the sequence for Humans would be titled, "Animalia - Homo sapiens".

Next, you will want to align your sequences so that you can compare homologous regions or nucleotides. Align your sequences by first highlighting all of your sequences (command +A) and then go to "alignment," (you may need to open your file at this step), and choose "Align by ClustalW" in the menu at the top of the page (or click the W button). Just use the default settings and click "OK." Then click off of your sequences to see the colors again. What did this do?

The Alignment Explorer is used to align all of your sequences (so that you are comparing the same position in each sequence). Why is sequence alignment so important? Because we need to compare homologous amino acid positions in the sequence! For example… Let's imagine there is an amino acid insertion in one of your sequences. If there was a three amino acid insertion (HQQ) starting at position 2 in the sequence of species D, then you will learn nothing from comparing your sequences as is; all of the remaining homologous amino acids will be shifted three positions later in species D relative to the other sequences:

```
Position      1  2  3 4  5  6  7 8  9 10 11 12 13
Species A     A L C A G A E A P  G  Q  Q  C
Species B     A L C A G A E A P  G  Q  Q  C
Species C     A L C A G A E A P  G  Q  Q  C
Species D     A H Q Q L C A G A E A  P  G
                  ↑ ↑ ↑
```

Thus, if we failed to align the sequences, even though there would only be a 3 amino acid difference between species D and the other species (due to the insertion), we would detect 12 differences between the sequences (at positions 2-13), and we would not recognize the string of 9 identical nucleotides that followed the insertion (underlined).

By aligning the sequences we would instead get:

```
Position      1  2  3 4  5  6  7  8  9 10 11 12 13
Species A     A - - -  L C A G  A E  A  P  G
Species B     A - - -  L C A G  A E  A  P  G
Species C     A - - -  L C A G  A E  A  P  G
Species D     A H Q Q L C A G A E  A  P  G
```

By properly aligning the sequences, we can see that following the insertion, the next 9 amino acids are identical in all 4 sequences. Thus we realize that species D is much more closely related to the other 3 species than it would have appeared had we not aligned the sequences.

Why is it important to align the amino acid sequences before we use them to build a phylogeny?
http://www.hhmi.org/biointeractive/creating-phylogenetic-trees-dna-sequences
(this link has slides… 9-16 that illustrate this, visit for fun if you wish)

8. After aligning your sequences, save the alignment!

Click Data at the very top of the alignment screen, then Export Alignment. Once there, select "MEGA Format" and save the file.

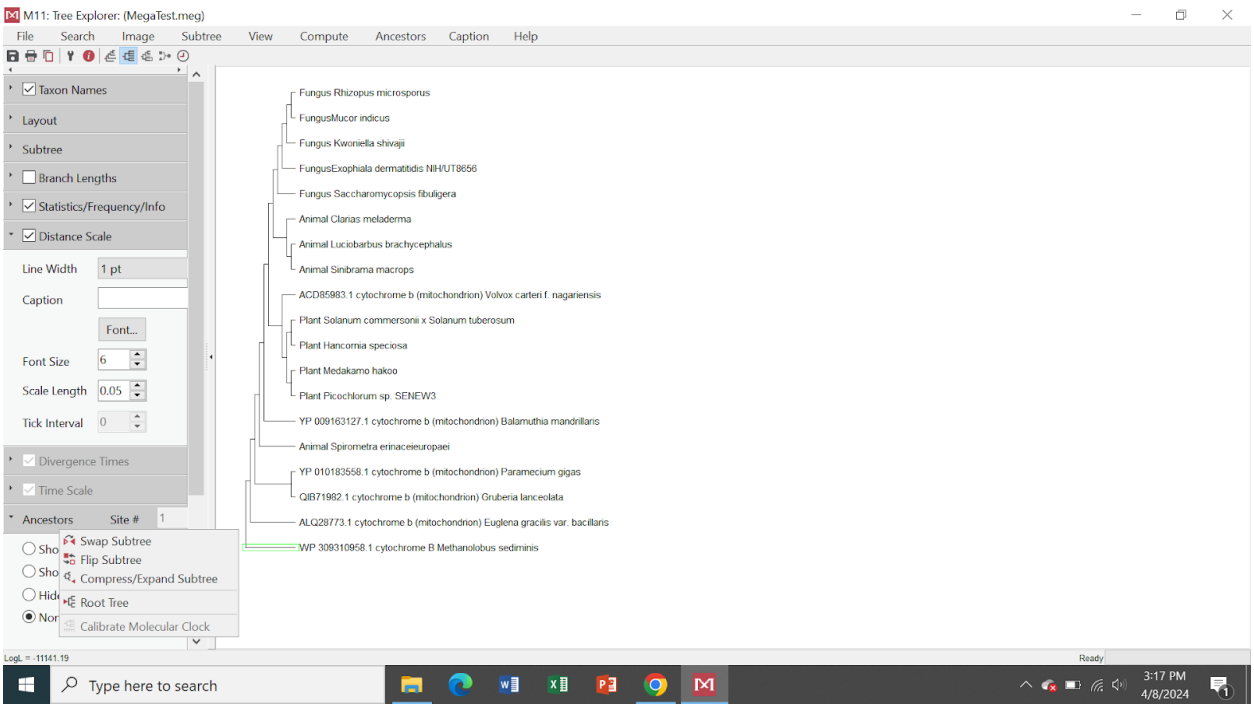
9. Once you have saved the alignment, go to the Original menu that you started with and click "Phylogeny" (the picture of the evolutionary tree) and then click Construct/Test Maximum Likelihood Tree.

Under "tests of phylogeny," click "Bootstrap method."

You may choose to change the "No. of Bootstrap Replications" to 100 rather than 500, since 500 may take awhile. For many Macs, bootstrapping is a glacially slow process. We recommend using one of the lab's PC laptops if your Mac bootstrapping is taking too long.

After finishing the bootstrapping process, you will be presented with two trees: the "Original Tree" and "Bootstrap Consensus Tree". We're going to stick with the "Original Tree" for now. The numbers you can see on the trees are the confidence that the program has in the tree its formed. The large the number, the more confident it is in that relationship.

10. On the "Original Tree", you will want to root the tree around your outgroup. Right click on the line coming out from the outgroup, then click "Root Tree". This will rearrange the tree to have the outgroup as the most distantly related individual.



11. SAVING. You can save your tree once it is built—again this is a good idea in case there is a crash. You can edit the tree, including changing the names (if you didn't in the previous steps). Once you are satisfied, go to Image>Save as PNG (or other format you prefer) or take a screen shot. Save these in a place you will be able to access. If you are using a lab computer, files will save onto that specific computer, so you should save to google drive or email your file to yourself.

Look at your tree. It should be satisfying!! You took protein sequences that researchers obtained in a lab by taking tissue from an organism, extracting DNA or protein, and sequencing it, and you turned them into a tree, unlocking millions of years of evolutionary history of these species!!

12. Analyze your tree…

How does the tree you created compare to the tree you hypothesized? Do each of the major Kingdoms (animals, plants, fungi, protists) form a monophyletic group in your tree? Would you expect them to form monophyletic trees?