

## Introduction:

Like many other scientists, biologists have to rely on computers to solve various problems. One type of problem that requires the assistance of a computer is comparing proteins from different species. In the lab that you're about to do, you'll learn to use some newly developed software that professional biologists use on a daily basis to compare proteins.

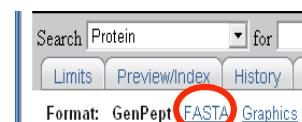
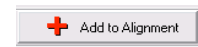
*Tip for the curious:*

You'll see the names of many kinds of animals during this investigation. You may find it helpful to keep a web browser open to Wikipedia, so you can see pictures and brief descriptions of these animals as you go.



## HOW TO RETRIEVE SEQUENCES OVER THE INTERNET USING MEGA 4

1. Open **MEGA 4**
2. From the "Alignment" menu at the top of the MEGA 4 window, select "Open Saved Alignment Session."
3. In the pop-up window locate and open the file named "Project\_MEGA\_4."
4. The Alignment Explorer window will open. Minimize this window for now and go back to the main MEGA window.
5. From the "Alignment" menu at the top of the MEGA 4 window, select "Query Databanks."
6. The **NCBI Entrez** webpage will automatically open in a separate window.
7. Next to the "Search" prompt, select "Protein" near the top of the list.
8. Next to the "for" prompt, type "**PAXNEB Pan Troglodytes**" and click "Go" (**btw, PAXNEB is a gene that controls eye development**).
9. You will be taken to a **flat file** for the protein sequence.
10. Find the source organism for this protein sequence and record it on your sheet. It will be listed near the top of the flat file next to the word source.
11. Scroll down to the bottom of this file and you'll see a 424 amino acid sequence (*note: each letter represents a different amino acid*).
12. To add this sequence to your alignment click the "Add to Alignment" button at the top of the window as shown here.
13. A popup window should confirm that the sequence was added successfully.
14. Next, you will use the protein sequence you just obtained to find similar protein sequences by performing a **BLAST** search.
15. Go to the top of the flat file and next to "Format" click "**FASTA**" as shown to the right.
16. The displayed sequence will begin with a header that starts with a ">" symbol and includes the name and **accession number**.
17. Highlight the sequence including the header and then copy it (Ctrl-C).



18. Click the back arrow twice (or until you get back to the NCBI Entrez webpage).
19. From the left side of the Entrez webpage, select “BLAST” (under “Related Resources”).
20. Just below “Basic BLAST” select “protein blast.”
21. Now paste (Ctrl-V) your sequence into the window just below “Enter Query Sequence.”
22. Click the blue “BLAST” button.
23. It may take several moments, but eventually a list of results will appear.
  - a. Note: A window may pop up with an error message. Click “OK.” This will not cause any problems, but may pop up frequently.
24. Scroll down to the box titled “Descriptions.” Try to find PAXNEB genes from all of the organisms on your handout, and then add them to the Alignment Explorer.
  - a. Tip: Use the find function to locate these organisms faster. You can do this by hitting the Ctrl + f keys at the same time and typing part of the organism’s scientific name into the box that pops up.
25. Click on the accession number (in blue in the leftmost column) and you will be taken to a flat file for the sequence.
26. Scroll down to the bottom of the report and observe the protein sequence. Add this sequence to the Alignment Explorer by clicking the “Add to Alignment” button at the top of the screen. A box will pop up to let you know the sequence was successfully added. Click “OK.”
27. Copy down the scientific name and the common name of the organism on your worksheet. The organism’s common and scientific names can be found at the top of the report to the right of “SOURCE.”
28. Click the back button on the browser and repeat steps 21 through 24 for the other organisms you wish to add to the Alignment Explorer. You’ll only need one sequence from each organism.
29. Once you have added all the sequences you want you may close the NCBI Sequence Viewer window.

#### ALIGNING SELECTED SEQUENCES USING THE ALIGNMENT EXPLORER

30. In the Alignment Explorer window, select “Save Session” from the Data menu and name your file.
31. From the Edit menu, select “Select All.” All of the sequences will be highlighted in blue.
32. From the Alignment menu, select “Align by **ClustalW**.”
33. Under Pairwise Alignment in the Clustal Parameters window change the gap opening penalty from 10 to 35 and the gap extension penalty from 0.1 to 0.75.
34. Under Multiple Alignment change the gap opening penalty from 10 to 15 and the gap extension penalty from 0.2 to 0.3.
35. Click “OK” and **MEGA 4** will align the sequences you selected using the **ClustalW alignment algorithm**. (Alignment places highly similar, conserved regions in vertical columns by inserting gaps between and around them. *Note: the gaps show up as hyphens.*)
36. In the **Alignment Explorer**, select any random letter or dash in the display window. This will cause all other amino acids in the alignment to adopt colors that reflect their biochemical properties.
37. Scroll through the alignment and note how strongly similar (conserved) regions have been placed into vertical alignment. (The gaps were inserted to account for length differences among the sequences.)

38. Sometimes regions at the beginning or end of the alignment will be poorly aligned, because the sequences were too dissimilar in length. We should crop off these poorly aligned regions on the ends of the alignment before moving on.
39. To remove these areas, first go to the far left side of the alignment.
40. Look for a small gray box on the top of the first column in the alignment and select it.
41. Now scroll to the right until you encounter a large conserved block, which will look like several columns with few gaps in them.
42. Find the leftmost column in this block. Now hold down the shift key and select the little gray box above the column just to the left of this place.
43. This will select all of the amino acids in the not-so-well aligned region from the beginning up to this point.




44. From the Edit menu, select “Delete” and Poof! They’re gone.
45. Repeat this procedure for both ends of the alignment, if needed. (You’ll need to select and scroll starting from the opposite end to do this.)
46. Now it is time to export the trimmed, aligned sequences as a **MEGA file** (with a “.meg” extension). From the Data menu, select “Export Alignment → MEGA Format” and name the file. You will also be prompted to name your input data.
47. You are now finished with the **Alignment Explorer** and may close it.
48. When you close the **Alignment Explorer**, you will be asked if you would like to open the saved MEGA file. Select “Yes.”

## INVESTIGATING THE ALIGNED DATA WITH THE **SEQUENCE DATA EXPLORER**

49. When the MEGA file is opened, a new window will also be opened. This is the **Sequence Data Explorer** window; we’ll use it to examine the aligned data.
50. The data in the **Sequence Data Explorer** is shown in black and white. To change it to a color-coded format, select “Color Cells” from the Display menu.
51. You will notice that there are amino acids, in the single letter code, running along the top row of the **Sequence Data Explorer**; these amino acids are known as the **reference sequence**.
52. Below this each cell will contain: (1) a “•” for a match to the reference sequence; (2) a letter, when an amino acid does not match the reference sequence; or (3) a “-” wherever a gap has been inserted.
53. What we’ve done up to this point is to make an alignment of PAXNEB genes from many different animals.
54. Did you notice how some animals have PAXNEB genes that are more similar and others have PAXNEB genes that have more differences? This information will be used to do the next step...

## INFERRING A PHYLOGENETIC TREE USING THE MAIN **MEGA 4** WINDOW

55. Now go to the main **MEGA 4** window.
  56. From the Phylogeny menu, select “Construct Phylogeny → **Maximum Parsimony (MP)**...”
  57. In the new window that opens, click on the tab that says “Test of phylogeny.”
  58. Select “**Bootstrap**” and accept the default number of Replicates (which should be 500), then click on the red check mark.
  59. Now select the button on the bottom of the Analysis Preferences box that says “Compute” and has a green check mark on it.
  60. In a few moments, you will have an inferred phylogeny, based on your aligned data set. This tree shape implies the fewest overall amino acid substitutions for your aligned protein sequences.
  61. Root your tree by clicking on the rooting button (shown to the right) on the top of the left hand toolbar under the arrow and then clicking on the branch leading to the sea anemone. The sea anemone should now be at the bottom of the tree.
- 
62. You can save a picture of the tree as follows: In the **Tree Explorer**, from the Image Menu, select “Save as TIFF file.”
  63. Print this file and label where each of the following characteristics first appears on the tree.
    - a. Characteristics: body hair, amniotic egg, forelimbs and hindlimbs, hinged jaw, large brain, mammary gland, opposable thumb, left-right symmetry, placenta, mesoderm, nipples, and vertebrae.
  64. You have just created a very powerful tool for exploring the history and diversification of a biological lineage!

## Student Handout

Find the PAXNEB protein sequence and common names for the following organisms. Use the blank spaces to fill in the common names for the five organisms that you add.

Scientific Name	Common Name
<i>Pan troglodytes</i> You added this when you gathered the BLAST sequence	
<i>Takifugu rubripes</i>	
<i>Bos Taurus</i>	
<i>Canis familiaris</i>	
<i>Rattus norvegicus</i>	
<i>Xenopus laevis</i>	