WALRUSES AND WHALES AND SEALS, OH MY!

Walruses and whales are both marine mammals. So are dolphins, seals, and manatee. They all have streamlined bodies, legs reduced to flippers, blubber under the skin and other adaptations for survival in the water. Although mammals evolved on land, these species have returned to the sea. Did they evolve from a single ancestor who returned to the ocean, or were there different return events and parallel evolution? We can't go back in time to observe what happened, but DNA sequences contain evidence about the relationships of living creatures. From these relationships, we can learn about the evolutionary history of marine mammals.

In this lab, we will use sequence information in GenBank (the public repository of all known DNA sequences from many species) and bioinformatics software called the Next Generation Biology Workbench (NGBW) to test hypotheses about the relationship between aquatic mammals (seals, whales, dolphins, walruses, manatees, and sea otters) and their potential ancestral relationship to land mammals.

We will use a protein that all mammals share: the hemoglobin beta protein. Hemoglobin is a good test molecule since it shows both conservation across species (since it performs the essential function of carrying oxygen in the blood), and variation between species. Species with unique challenges, such as holding their breath for long underwater dives, may have evolved changes in their hemoglobin which improved their supply of oxygen. In addition, hemoglobin has been studied by many evolutionary biologists, so sequences are available in GenBank from many different organisms.

In this lab, we will be testing hypotheses about the evolutionary ancestry of different marine mammals. To repeat, we are trying to answer the question: **Did marine mammals evolve from a single ancestor who returned to the ocean, or were there distinct return events from separate ancestors?** As a starting point, *let's hypothesize that marine mammals have a single common land mammal ancestor*.

In this exercise, we will compare the hemoglobin proteins of these marine mammals to some representatives of the major taxa of land mammals.

MARINE MAMMALS:	**LAND MAMMALS:**
minke whale (baleen whales)	carnivora: dog, *Canis familiaris*
dolphin (toothed whales)	rodentia: rat, *Rattus norvegicus*
harbor seal	herbivore: cow, *Bos taurus*
walrus	primates: human *Homo sapiens*
otter	proboscidea: African elephant, *Loxodonta africana*
manatee	marsupials: red kangaroo, *Macropus rufus*

To proceed through this exercise, first visit the Next Generation Biology Workbench (http://www.ngbw.org), and create an account. To do this, click on the "Use the Workbench Now" button. A new page will open. On that page, click on the button that says "Create an account." This will open a form in which you enter your personal information. Once your registration is complete, you will be logged in to your personal area.

The NGBW allows you to store data and tasks in folders just like MS Outlook or other mail clients. So before working here, you must create at least one folder. Click on the button that says "Create New Folder". Name your folder "IB BIOLOGY LAB", and save the name. The folder you created will appear on the left side of the screen. Open a second browser page or new tab so you can toggle between the NGBW and the sequence database we use. Now, we can begin the exercise.

PART A: FINDING AMINO ACID SEQUENCES

1. First we need to get the sequence data for the hemoglobin protein from marine and land animals. Several sites offer this kind of data. For example, **UniProt**, is protein sequence database hosted by the EBI (http://www.uniprot.org/)

2. At the top of the UniProt home page, there is a search tool. Make sure the drop down under ***Search in*** says "Protein Knowledgebase: UniProtKB." You will search for your data here. Just as with any Google search, effective searching for data in a protein database depends on choosing effective search terms. One thing you need to know: the hemoglobin beta gene and protein is known as "**HBB**" in UniProt and other databases. However, there are many, many hemoglobin sequences available, we need to be as specific as possible about the identity of the sequences we want. That way we won't be overwhelmed by search results we are not interested in.

We want **to gather sequences from the 12 species listed above**. To retrieve these sequences, type in the query box: HBB and the name of the organism you want. For example, type "HBB minke whale" and click "Search".

NOTE: This database is usually queried by computers, so it does not always return results that seem organized logically. Don't get frustrated; be patient. Your search will often return several answers. Look at the results for a record name that contains the terms Hemoglobin, and Beta. That will be the correct protein! You can insure this by including these as search terms. Because short strings like "cow" and "dog" are fairly non-specific and occur frequently in databases, you should use the scientific names *Bos taurus* and *Canis familiaris* as search terms.

3. Find the sequence you need among the returned sequences. Now save it for further analysis using the NGBW. To do this, click the Accession number hyperlink of the sequence you want. A page will open with a wide variety of information about the sequence. Today, we just want to recover the protein sequence alone. To do this, click the gold FASTA link at the extreme right of the sequence page (its small, but its there!). This opens a file containing the amino acid sequence of the protein, represented using single letters for each of the 20 naturally occurring amino acids (A=alanine, M=methionine, P=proline, etc). Here is how the file looks:

```
>sp|P02107|HBB_MACRU Hemoglobin subunit beta OS=Macropus rufus GN=HBB PE=1 SV=1
VHLTAEEKNAITSLWGKVAIEQTGGEALGRLLIVYPWTSRFFDHFGDLSNAKAVMGNPKV
LAHGAKVLVAFGDAIKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNIIVICLAEHFGK
EFTIDTQVAWQKLVAGVANALAHKYH
```

While this file isn't nice for humans to read, the Biology Workbench knows how to access and utilize this sequence information for your analysis. The first line is called a "header". It contains information describing the sequence. The reminder represents the sequence of amino acids in the protein. Right click on the sequence, and "copy" it.

4. Transfer the sequence to the NGBW by going to the tab or browser page with your open NGBW account, and click the "Data" folder. Click the link to upload a file. This opens a data management area. In the "Label" entry box, give the sequence a name you will recognize easily (for example, the common name of the organism). Use the right click, and paste the data into the form box labeled "***Or, enter your data:***" Paste your data into this block.

Before saving it, let's change the **sequence header** at the top of the protein sequence that we pasted into the box. This will help us later. The header is changed to be more informative when we make a tree later. This header can be edited for clarity, **but you MUST preserve the ">" symbol**. You can use the scientific name or the common name to identify your sequence. For example: The red kangaroo sequence begins like this:

```
>sp|P02107|HBB_MACRU Hemoglobin subunit beta OS=Macropus rufus GN=HBB PE=1 SV=1
```

This can be edited to simply say this: >Red_kangaroo

WARNING: ***do not introduce blank space into the header line***. For example, type >Red_kangaroo or >Redkangaroo, BUT NOT >Red kangaroo

Do not be concerned with blanks spaces or line breaks in the sequence itself. These will be ignored.

Use the dropdown boxes at the bottom of the page to tell the application what kind of data you are uploading. The Entity Type is "Protein", the Data Type is "Sequence", and the format is "FASTA". Now select "SAVE". You only need to set these once, the application will remember after that.

5. Return to UniProt (on the other web page). Copy and Paste the remaining hemoglobin protein sequences into the NGBW in exactly the same way. You should have 12 data items in your data folder when you are finished.

ALIGN YOUR SEQUENCES

Sequence alignment is a tool evolutionary biologists use to assess how closely related two sequences are. One tool for this is **ClustalW**, a software program that aligns protein and DNA sequences.

1. To use **ClustalW** (and other software packages) in the NGBW, you create "**Tasks**." Here's how. Click on the "**Tasks**" icon in your working folder. When the Task management page appears, click on the "**Create New Task**" Button. This will open a task creation window.

2. Enter a description for the task you will recognize in case you need to come back later. "SEQUENCE ALIGNMENT" might be a good description. Type that into the box and click the "**Set Description**" button.

3. Now click on the "**Select Input Data**" button. Check the click boxes to the left of all the sequences you have entered for this exercise. Click the "**Add selected to task**" button.

4. When the Task Creation Pane re-appears, click the "**Select Tool**" button. From the **Protein Sequence Tools tab**, choose "**CLUSTALW_P**" (the tools are alphabetical). Now click the "**Save and Run Task**" buttons.

5. Be patient, and eventually a new page will load that lets you follow the progress of your jobs. Click the "**Refresh Tasks**" tab near the top of the page, until the "**View Status**" button on the right turns into "**View Output**." *This may take a few minutes!* **Click on the "View Output**" tab, and a page showing your results will appear. Click on the "**View**" button for "**outfile.aln**".

6. Examine the output file. It shows the calculated alignment, and the changes in amino acid identity at each hemoglobin Beta position as the species evolve. Save the alignment data, by clicking the "**Save to Current Folder**" button. Use the drop down boxes to set the Entity type to "**Protein**", the Data type to "**Sequence Alignmen**t", and the Format Type to "**FASTA**." Now click "**Save**".

   NOTE THE FOLLOWING **WARNING BEFORE DOING THE NEXT STEP**: One thing that has killed the server in the past is when students submit fasta files to Boxshade instead of the alignment (.aln) files. If 8 people make this mistake, it can bring the server to a halt. This problem is a bug in Boxshade, be extra careful about not submitting fasta files for boxshade. Make sure you follow the tutorial step by step and avoid bringing down the server. If you notice the server is down; use the contact info to notify NGWB. Please be careful and don't cause unnecessary frustration for your fellow students.

7. It is easier to see which parts of the protein are well-conserved (unchanged) and which parts of the protein have experienced mutations by preparing a color coded diagram to emphasize the changes. To do this, you can create a task that uses the program **Boxshade**. Click the tasks link listed under your folder on the left side of the page. Click the "**Create New Task**" button. Give the task a description ("boxshade" would be a good name) and **"Set the Description**", just like before. Click the "**Select Input**" button, and find your alignment data, check the box to the left of the alignment you created above (the .aln file), and click "**Add Selected to Task**." When the task creation page reloads, choose Select Tool, and find "**BOXSHADE**" under the **Phylogeny/Alignment Tools tab**. *This page might take a minute to load.* Then click the "**Save**" button. When the task creation page re-appears, click the "**Save and Run**" button.

8. As you did before, click the "refresh tasks" button until the "view status button" changes to "view output" for the boxshade tool. To view a color coded alignment, click on the "View **Output**" button of this task, and click the "**View**" button for the file **boxshade.html**. The portions of the sequence that are conserved between all species will be highlighted in black. Amino acids that are similar, but not identical will be shown with a gray background. Those that have a different character will be shown with a white background. These represent areas of the amino acid sequence that have experienced (and tolerated) genetic changes.

   Take a moment and look at the screen in front of you. Each line is the amino acid sequence of the same protein (hemoglobin beta) in different species. It is interesting to scan along the amino acid sequences and look how they line up — how they are the same in the different species and how they are different. You can see the traces of evolutionary processes here: where amino acids have changed where they have stayed the same, and where amino acids have been lost. You are looking at the record of evolutionary history!

9. Take a screenshot of this alignment chart to use in your lab report (press the "Printscreen" key, typically labeled "PrtScn"). The picture of the screen is now waiting to be pasted into a document. Paste the screenshot into a Word or Open Office document to use later.

DETERMINE THE DISTANCE BETWEEN SEQUENCE PAIRS

1. While your eye can tell qualitatively what has occurred between species, rigorous analysis requires that we quantify these differences. The next step is to quantify the evolutionary distance between sequence pairs. To do this, go back to the Tasks area of your folder. Click the **"Create New Task**" button. Give the task a description ("DISTANCE" would be a good description) and **"Set the Description"**, just like before. Click the "**Select Data**" button, and find your alignment data, check the box to the left of the alignment you created above (the .aln file), and click "**Select Data**."

2. When the task creation page reloads, choose **Select Too**l, and find "**CLUSTALW_DIST"** under the **Phylogeny/Alignment Tools** tab. *This page might take a minute to load.* When the task creation page re-appears, click the **"Save and Run"** button.

3. When the Task management page reloads, use the "**Refresh Tasks**" button to monitor when the job completes. When the "**View Output**" button appears, click on it, and expose the results. Click on the **View** button for the **"infile.dst"** link to expose the Distance Matrix results. Create a data table to record the distance matrix data.

4. The CLUSTALW_DIST program also produces a phylogenetic tree inferred from this data. To view this tree return to the **Results** page, and click the "**View**" button for the file "**infile.ph**" You will see file containing the names of your sequences, separated by colons and parentheses. You can't interpret this result directly, but the NGBW provides an application to do that for you. Click the "**Save to Current Folder**" button. Use the drop down boxes to set the **Entity type to "Taxon**", the **Data type to "Phylogenetic Tree**", and the **Format Type to "Newick**." Now click "**Save**".

VIEW THE TREE

1. To view the tree in a human-interpretable form, **return to the Data area of your folder,** and click the tab "**Phylogenetic Trees**". Click the the infile.ph file you just created and it will open up, revealing these two links "Show/Hide Data Contents | Draw Tree." Click on the "**Draw Tree**" link, and you will see an interactive view of the Tree.

2. In the interpretation of the phylogenetic tree, it is useful to "root" the tree using a more distant relative of the add one more step to make our tree more accurate. We need to add an "**outgroup**" to the mix of species we are analyzing. An **outgroup** clarifies the evolutionary relationships by providing a "root" to the tree. It is designated as a species that is far removed from any of the other species we are comparing. Here, the red kangaroo was included for this purpose, since it is a marsupial in contrast to all the other mammals in our study, which are all placentals. Therefore, the kangaroo is selected to be the *most different* organism from the other mammals on your tree. Redraw the tree with red kangaroo as the outgroup, by entering the name you entered after the > mark on the header file of your fasta sequence corresponding to red kangaroo. If you followed this exercise exactly as above, **enter red_kangaroo, and click "Redraw**". If you get the message "None of the outgroup taxa are valid" it means you mistyped the outgroup name, or you forgot the name you chose. The case of the characters is important, and must also match exactly. Record a screen shot of the phylogenic tree for your lab report.

   NOTE: Phylogenetic trees built with this software can only be used to make conclusions about common ancestry. They cannot be used to make conclusions about the timeframe of evolution. The length of branches is not a measure of evolutionary time. It is merely an artifact of physically arranging the tree.

LAB REQUIREMENTS

Problem question
Hypothesis
Screen shots (alignment, distance matrix, and phylogenic tree)
Conclusion

- Did your tree support your hypothesis? Explain.
- What does this phylogenetic tree structure suggest about the evolutionary history of marine mammals? Go into detail here about what parts of the tree lead you to what conclusions about the evolutionary history of the marine mammals.
- If marine mammals share common morphological characteristics, what do your conclusions about their evolutionary history imply about these common characteristics?

Questions

- What is the advantage of using the protein sequence from the hemoglobin beta gene to prepare the comparisons between species?
- What organism served as your outgroup? Why? What function does the outgroup serve?