In this lab you will be guided through the process of constructing two phylogenetic trees each constructed using a different method; maximum parsimony and maximum likelihood. You will be using several programs implemented in the PHYLIP software package (<http://evolution.genetics.washington.edu/phylip.html>).

PHYLIP Overview

The PHYLogeny Inference Package (PHYLIP) is composed of 35 different programs. These programs can be divided into several different categories: phylogeny construction (including distance and maximum likelihood methods), tree drawing, editing and comparison. Detailed descriptions about the programs in each category are available (http://evolution.genetics.washington.edu/phylip/phylip.html).

**PHYLIP download and installation:**

Detailed instructions describing how to fetch and install the PHYLIP software package on your computer are available at the following websites (<http://evolution.genetics.washington.edu/phylip/getme.html>, and <http://evolution.genetics.washington.edu/phylip/install.html>), respectively.

 This lab will focus on the use of some of these programs in order to construct a phylogeny that depicts the putative evolutionary relationship between several proteins. We will be using two different phylogeny construction programs: “protpars” (protein parsimony implementation) and “proml” (protein maximum likelihood implementation). From the Tree drawing program category we will use the “consense” and “drawgram” programs to build the consensus tree and visualize our trees, respectively.

1. Collecting Sequence Data

The construction of our phylogenetic trees requires, as an initial step, the collection and alignment of several protein sequences. We will use the myoglobin protein as our example. We will need to gather myoglobin sequences from several different species. Below are step by step instructions for collecting the data:

1. Open a browser window and go to the NCBI website (<http://www.ncbi.nlm.nih.gov/>).
2. In the Search drop down menu select “Protein” and enter “Q8T7J9” in the text box then click the **Go** button.
3. Select the Q8T7J9 link to open a page providing detailed information about the specified globin. Click on the button next to the “Display Settings” link and select the FASTA radio button and then click “Apply” (see image below). The web page will refresh and display the sequence in FASTA format.



Copy this sequence to a text file.

Question 1: What is the scientific name (Genus & species) of the organism this globin protein came from?

1. Rename the sequence in the text file by replacing the original FASTA sequence name line (i.e. everything after the > symbol) with the suggested name provided in the table 1 below.
2. Repeat steps 2 -4 for each myoglobin/globin sequence in table 1. **Add all sequences to the same text file**.

Question 2: As you retrieve the sequences fill in the blanks in Table 1 with the common name of the organism in which the corresponding protein was found.

Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Sequence Number | Protein ID | Suggested Name | Common name |
| 1 | Q8T7J9 | Yeightsi | Mollusk |
| 2 | NP\_976312 | Hsapien1 | Human |
| 3 | NP\_976311 | Hsapien2 | Human |
| 4 | NP\_005359 | Hsapien3 | Human |
| 5 | NP\_038621 | Mmusculus | Mouse |
| 6 | NP\_067599 | Rnorvegicu | Rat |
| 7 | NP\_999401 | Sscrofa | Pig |
| 8 | NP\_776306 | Btaurus | Cattle |
| 9 | ABN71515 | Iiguana | Common Iguana |
| 10 | P02185 | Pcatodon |  |
| 11 | P68082 | Ecaballus | Horse |
| 12 | P02205 | Talbacares |  |
| 13 | P02155 | Ssciureus |  |
| 14 | P11343 | Llutra | Eurasian river otter |
| 15 | P02197 | Ggallus | Chicken |
| 16 | P56208 | Ccaretta |  |

1. Align the protein sequences using CLUSTALW (see lab 3)(<http://align.genome.jp/>). Change the output format to “Phylip” and execute the multiple sequence alignment. Once the alignment output page has loaded save the alignment by scrolling down the page and right-clicking on the ‘clustalw.phy’ link on the clustalw alignment output and then clicking on the ‘**Save Target As…**’button. Save the clustalw.phy file to the Phylip\exe folder which should be on the local disk C; the directory path to the Phylip exe directory should be Local disk C:\phylip-3.69\exe.

The Phylip format is a common multiple alignment format. There are two styles; interleaves and sequential. The CLUSTALW output will be interleaved. The first line of the file will have two numbers the first indicating the number of sequences aligned and the second showing the length of the multiple sequence alignment. Notice that there is limited space for the species names at the beginning of the sequence lines (Phylip format limits the species names to 10 characters, which is why you condensed the names in your FASTA file; be careful that no two sequences in your FASTA file share the exact same first 10 characters or your will have sequences with the same name in your phylip format output).

E.g. of interleaved Phylip format:



1. Construct Maximum Parsimony Tree

Now we will create a phylogenetic tree of the 16 myoglobin proteins using the “protpars” program. The PHYLIP software package is in a folder named “phylip-3.69” on your desktop. This program counts the number of non-synonymous nucleotide changes (single nucleotide changes which alter the amino acid), and finds the tree which includes all the input sequences and requires the fewest number of changes.

1. Open the “exe” subdirectory inside the “phylip-3.68” folder. This folder contains all the PHYLIP package executables.
2. Make sure that your alignment file is present in this folder.
3. Double click on the “protpars” icon:
4. Enter your alignment file name (i.e. whatever you named the clustalw.phy file) and press <Enter>:



1. Here is a screen shot of the “protpars” options menu:



If you are interested in the full description of each option see the website (<http://evolution.genetics.washington.edu/phylip/doc/main.html>).

1. Change the “Randomize input order of sequences” by pressing ‘J’ then <ENTER>
2. You will be asked to enter a random odd number, enter “5” then <ENTER>
3. You will be asked for the number of times to jumble, enter “5” then <ENTER>
4. Notice that the menu entry for Randomize input order had changed:



 The randomization of the input order of sequences will ensure that the order of adding the species to the tree does not bias the end tree topology. (You may want to randomize the data more in practice, but to save time during the lab we will keep the randomization fast).

1. Change the “Outgroup root” by pressing ‘O’ then <ENTER>
2. You will be asked to enter the number of the outgroup, enter “16” then <ENTER>

Species 16 should be the mollusk species “Yeightsi”.

1. Execute the program by entering ‘y’ then <ENTER>

The program will then create two files: outfile and outtree. The “outtree” file contains a set of 3 trees in Newick format which are all equally parsimonious. The “outfile” contains all the parsimony statistics and and rough graphical representations of each of parsimonious tree.

Question 3.

Baesed on the lecture notes, how many different **rooted** trees are possible with 16 species?

1. Rename the outfile and outtree files to something informative. Almost all the Phylip programs will create output files named outfile or outtree and if you are not careful your results from other programs may be overwritten and lost.
2. If your outtree file has more than one parsimonious tree then the next step is to create a consensus tree of the 3 most parsimonious tree using the “consense“ program. Run “consense” by double clicking on its icon:
3. Enter the outtree filename:



1. Screen capture of “Consense” options menu



We are building a rooted tree because we are assuming that the Mollusk *Y. eightsi* is the most ancient animal species in the tree. Change option “Trees to be treated as Rooted” to ‘yes’ by typing ‘r’ then <ENTER>

1. Execute the “Consense” program by typing ‘y’, then <ENTER>

The outtree file will contain the putative consensus tree in newick format and the outfile contains detailed information about the consensus tree analysis.

1. Rename the outtree and outfile files something meaningful.

1. Construct Maximum Likelihood Tree

 Using the same myoglobin multiple alignment file used to construct the parsimony tree we will construct a maximum likelihood tree. This method has been implemented in the “proml” program.

1. Start the maximum likelihood program by double-clicking on the “proml” icon:
2. Enter the multiple alignment file name:



1. Screen capture of “proml” options menu: As with the “protpars” program, we will not be changing many of these options, a full description of each is available from the website (<http://evolution.genetics.washington.edu/phylip/doc/main.html>)
2. Change the ‘Speedier but rougher analysis’ option to ‘No’ by typing ‘S’ then <ENTER>
3. Change the ‘Randomize input order of sequences’ to ‘Yes’ with a seed of 5 and jumble 5 times (see instructions 6, 7, and 8 for the protpars program).
4. Change the ‘Global rearrangments’ option to ‘Yes’ by typing ‘G’ then <ENTER>.
5. Change the ‘Outgroup root’ option to ‘Yes’ and specify sequence 16 as the outgroup (see instructions 10 and 11 for the protpars program).
6. Execute “proml” by typing ‘y’ then <ENTER>

Depending on the speed of your computer this program may take several minutes to run.

Viewing your Trees

 There are numerous programs available that will display your trees such as “drawgram” which we will be using in the lab. Other programs include phylowidget (<http://www.phylowidget.org/>) and iTOL (interactive tree of life) (<http://itol.embl.de/>) which are online tree visualization tools.



1. Start “drawgram” by double clicking on the icon:
2. Enter the filename of your Maximum Parsimony Consensus tree:



1. You will then be asked to enter a font file name:



There should be font files in the /exe folder name “font1” to “font6”, you can enter any of the file names. E.g. type ‘font1’ then <ENTER>

1. Screen capture of “drawgram” options menu:



These options allow you to change the visual style of the tree from the overall size of the tree to the angle between the branches. We don’t need to change any options, but if you find that some of your labels at the end of your branches overlap then you may want to try changing some options. When you are ready, type ’y’ then <ENTER>.

A window will open displaying your tree.



Notice that all the species line up vertically, this is the default organization when the input tree file does not have branch lengths.

1. Repeat steps 1 – 4 using the maximum likelihood outtree file as input.

Notice that the species names do not line up in the maximum likelihood tree as they did in the maximum parsimony tree because branch lengths are automatically supplied in the likelihood tree.

Look at the maximum likelihood tree. The Iguana (Iiguana) and Loggerhead turtle (Ccaretta) shared a common ancestor (they branched from the same tree node). Notice that the branch leading to the turtle myoglobin is longer than the branch leading to the Iguana myoglobin.

Question 4. What does the difference in branch length indicate?

Compare the parsimony and likelihood tree and notice that there are several significant differences in the organization of the species in the trees. For example, the Llutra (Euroasian river otter) is included as an ungulate in the maximum parsimony tree whereas in the likelihood tree the river otter is more closely related to the rodents.

The trees also show several significant similarities.

Question 5. Based on the topology of the two trees provide three common species groupings found in both trees.

To get an idea of what the true phylogeny could be you can add more data (concatenate several gene/proteins together in the alignment), run a bootstrap analysis and try other tree construction methods (neighbor joining, Bayesian methods, etc).