Using Databases to Obtain Real Amino Acid Sequence Data to Compare the Myosin Heavy Chain in various Fish Species and Create Evolutionary Trees

Background:

In order to determine how closely related species are, scientists often will study amino acid sequences of essential proteins. Any difference in the amino acid sequence is noted and a phylogenetic tree is constructed based on the number of differences. More closely related species have fewer differences (i.e., they have more amino acid sequence in common) than more distantly related species.

There are many tools scientists can use to compare amino acid sequences of muscle protein. One such tool is the National Center for Biotechnology Information (NCBI) protein databases (http://www.ncbi.nlm.nih.gov/). By creating a file that contains the amino acid sequences of a protein you are interested in for various species, the *ClustalX* software compares the sequences to all others in the file and aligns them. The data generated provides enough information to construct cladograms.

ClustalX is a computer program that is used to do multiple sequence alignments using either DNA or protein. Once the program has made an alignment, C*lustalX* can draw phylogenetic trees connecting the sequences in the alignment. The phylogenetic tree must be opened with an appropriate viewing program such as *drawtree*, *phylip* toolkit, or *TreeView*. You will be using *TreeView* in this activity. *TreeView* is a simple program for displaying phylogenies on Apple Macintosh and Windows PCs.

Purpose:

The purpose of this activity is to use data obtained from NCBI along with the *ClustalX* and *TreeView* programs to construct an evolutionary tree based on the amino acid sequences of the myosin heavy chain. In this activity you will use the aquatic species we chose for the protein fingerprinting lab. You must first find the amino acid sequences of the myosin heavy chain for each species using NCBI. Then you will compare and align the sequences using *ClustalX*. This information can then be used by the program *TreeView* to automatically create a cladogram or evolutionary tree.

Materials:

- Use of a computer station, preferably one station per pair of students.
- Computers must have the following programs downloaded onto them:
  - *ClustalX*
  - *TreeView*

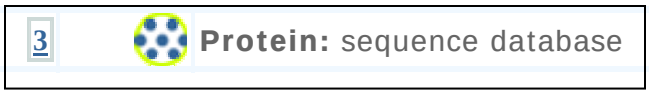
  *Both programs are available: http://www.cgal.icnet.uk/bioinformatics under downloads and then under PC). You will need WinZip to open the programs.
- Websites
  - NCBI Homepage http://www.ncbi.nlm.nih.gov/
  - NCBI Blast page http://www.ncbi.nlm.nih.gov/blast/
- File of FASTA sequences for myosin heavy chain.

Procedure:

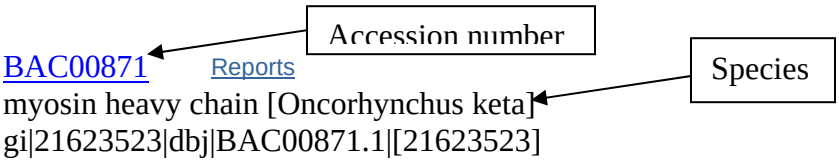
To find an amino acid sequence of interest:

1. Go to the NCBI website and click on “Entrez Home” in the menu on the right side of the page.

2. Where you see “Search across databases”, type in your query terms. For example, “myosin heavy chain” and salmon.
3. Entrez will locate all the information related to your query. For this activity you are interested in the amino acid sequences. Locate the following in the menu. As you can see there were three sequences located for myosin heavy chain and salmon.



4. Click on “Protein: sequence database”. This will bring you to a page that lists the accession number of the amino acid sequence along with the protein name and associated species.



5. Click on the accession number. This will bring you to a page that lists information including the publication material and the actual amino acid sequence. Scroll to the bottom where you can see the amino acid sequence.

   *Note: In these protein sequences, amino acids are represented by a one letter symbol much like nucleotides. Table 1 provides the abbreviations and symbols for the amino acids.

6. In order to compare this sequence in *ClustalX* with the other sequences you will gather, the sequence must be in the proper format called FASTA. Go back to the top of this page and locate “display”. Click on the drop down menu and select “FASTA”. Left-click and hold at the beginning of the sequence, in front of the carrot symbol. Drag and highlight until you reach the end of the sequence. Next right-click and *copy* the sequence.
7. Now go to “Start” at the bottom left corner of your screen and go into programs. Go into accessories and open “Notepad”. Here you will create a file to contain all of your sequences. Once in notepad you can right-click and *paste* the sequence.

   Make sure you save this file so that you can locate it later.

8. Now go back to Entrez and find your next sequence following the previous steps. Be sure to save the FASTA sequence in your notepad file. Skip a line between sequences.

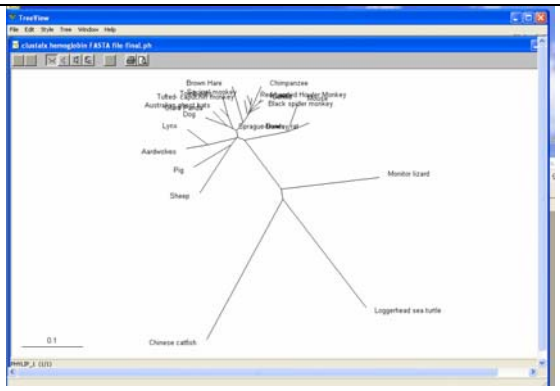
<u>To do a multiple sequence alignment using ClustalX and create a cladogram:</u>

1. Open the *ClustalX* program on your computer.
   a. Click on “Load Sequences” on the menu bar. Find and select the file that contains all of the protein sequences for the various species to be compared.
   b. Run the alignment by clicking on “Alignment” on the menu bar and select “do complete alignment”. Before the alignment begins, the program will ask you for two output file names. You can just click “Align” since you won’t be using these output files. The alignment may take a few minutes to complete—progress will be displayed at the bottom of the window.

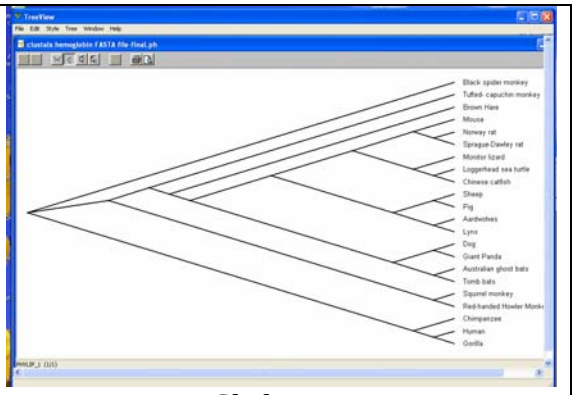
c. Once the alignment is complete, use the scroll bar to look at the overall alignment. Since the amino acid residues are highlighted with a different color, it's easy to see where the alignment is good. An asterisk (*) is printed at the top of the alignment where all sequences agree on a particular base location. A hyphen (-) is put in where a gap has been inserted by the program to make the alignment work well.

d. Before closing the program you must also save to disk the tree representation of this alignment.
   - Click on "Trees" on the menu bar.
   - Select Draw N-J Tree (this automatically saves it as a *.TRE* file)
   - You will not be able to visualize the phylogenetic tree in this program.

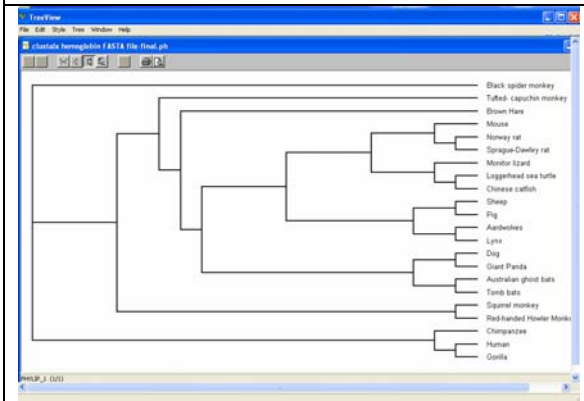
2. Now you will plot the tree arising from this alignment. Without closing the *ClustalX* program (just in case you need to go back to it), right-click on the *.TRE* file that you just saved, select "open with…" and choose the *TreeView* program. This will automatically open the file and display the phylogenetic tree.
   a. There are 4 different types of trees available. You can switch between them by using the buttons on the menu bar.



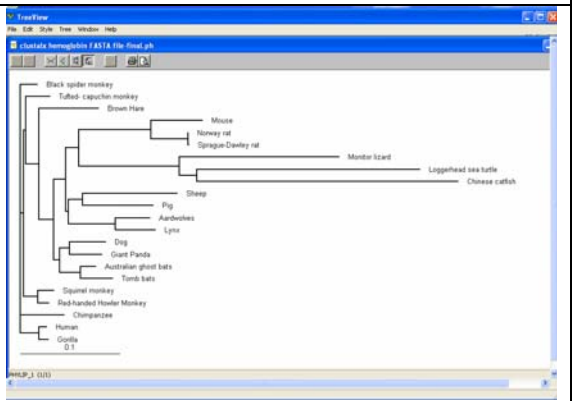
Unrooted tree or radial tree



Cladogram



Rectangular Cladogram



Phylogram

3. Compare your hand drawn phylogenetic tree and the cladogram you created from the protein fingerprint data with the computer generated tree.

Answer the following Focus Questions:

1. How is homology related to phylogeny?
2. How does the hand drawn phylogenetic tree compare with the computer generated model?
3. How does the cladogram you created from the protein fingerprint data compare with the computer generated model?
4. How can you tell how closely related 2 species are?
5. What might cause branching to occur in the tree?
6. Does the tree seem to reasonably represent the actual evolutionary distances between all selected organisms? (Look both at the big picture; i.e. the distances between the obviously very different organisms, as well as at individual pairs of organisms.) Explain.
7. What hypotheses would you propose to explain any unexpected features of the tree you produced?

Further Exploration:

Conduct your own investigation. Use the NCBI Entrez search page to find a protein related to a gene, disease, or biological process. You can explore the structure and function of any protein or study the evolutionary relatedness of different species based on that gene or protein. Make sure to identify a single gene or protein and record its accession number, GenBank, and FASTA format sequence files. When looking at homology, be sure to save the top 5-10 proteins homologous to your protein of interest. Obtain their FASTA format sequences by creating a text document with all FASTA reports listed one after the other. This could then be used by *ClustalX* for a multiple alignment.

Consider:

- What do you hypothesize is the function or enzymatic activity of your protein? Why?
- Are there specific portions of your protein that are present in other proteins? What function(s) might these shared protein domains serve?
- How might changes seen in the protein sequence between species reflect altered function of those proteins?

References:

Campbell, Reece, and Mitchell. (1999). *Biology*, 5th ed. by Benjamin Cummings.

Fiocruz Bioinformatics Training Course. (2005). "Molecular systematics and evolution: ClustalX Practical." [Online]. Available: http://www.dbbm.fiocruz.br/james/ClustalX_tutorial.html

The University of Arizona College of Agriculture and Life Sciences and the University of Arizona Library. (2005) "The Tree of Life Web Project." [Online] Available: http://tolweb.org/tree/phylogeny.html

Hershberger, R. (2000). "Identifying Homologous Sequences using BLAST Servers" on *The Bioactive Site*. [Online]. Available: http://www.rickhershberger.com/darwin2000/blast/1.