BIOLOGICAL FRAMEWORKS FOR ENGINEERS

Homework #3 (due 10/21/11) [NCBI, Blast, and Clustal]

The Internet has a wide range of free tools for manipulating and comparing DNA and protein sequences. Today we will use some of these tools to compare the protein sequences of one protein that is made by many organisms. The protein is actin, which is involved in muscle contractions in higher organisms like humans and in the motility of some single-celled organisms.

I. FINDING THE SEQUENCE OF ACTIN IN HUMANS

There are several ways of finding the sequence of proteins. The most convenient way is to go to NCBI, the National Center for Biotechnology Information, whose website is: www.ncbi.nlm.nih.gov

A. Since we are interested in protein sequence, choose "Proteins" on the left-hand side menu. Select "Protein" from the list of databases.

B. Now we will check on the information we have about a protein. The name of the protein is actin, and it is involved in muscle contractions. So type in "**actin AND muscle**" (case sensitive) and click Go.

C. You will get about 4218 hits from this search. On the right hand side, you will see a list of organisms with Homo Sapiens at the top. Click it.

D. This time you'll get about 863, which are all the actin muscle isoforms in humans. You will find that there are several kinds of muscle proteins located in different organs of human body.

Question 1: List at least 4 different kinds of muscle proteins, the number of amino acids, and their locus locator numbers (e.g. NP_001606.1).

G. Let's pick one to study further: BAA00546. In the Search menu, type "BAA00546" and click "go".

Question 2: Which chromosome is this protein's gene located on?

H. Find the amino acid sequence on the bottom of the page. Each letter represents an amino acid.

I. For your convenience, paste the sequence to a word document and give a name to the sequence.

For example:
>human enteric actin
1 mceeettalv cdngsglcka gfagddapra vfpsivgrpr hqgvmvgmgq kdsyvgdeaq
61 skrgiltlky piehgiitnw ddmekiwhhs fynelrvape ehptllteap lnpkanrekm
121 tqimfetfnv pamyvaiqav lslyasgrtt givldsgdgv thnvpiyegy alphaimrld
181 lagrdltdyl mkiltergys fvttaereiv rdikeklcyv aldfenemat aasssleks
241 yelpdgqvit ignerfrcpe tlfqpsfigm esagihetty nsimkcdidi rkdlyannvl
301 sggttmypgi adrmqkeita lapstmkiki iapperkysv wiggsilasl stfqqmwisk
361 peydeagpsi vhrkcf

II. LOOKING FOR SIMILAR SEQUENCES (BLAST)

Now that we have a protein sequence, we are going to use a program that looks for the closest sequences matching this protein in other organisms.

A. Go back to the NCBI home page. Choose "Blast" from the toolbar on the right.

B. Find the link "protein blast" under "Basic BLAST".

C. Copy your sequence with its name and, very importantly, with a ">" in front of the name; paste the whole thing to the "search" box. Click the blue "BLAST!" button. You might need to wait a minute or two until a new page comes out, and then click the blue button "Format!"

D. This will open a new page in another window containing the first 500 closest blast hits to your sequence.

Question 3: What is the range of the score (bits)?

Question 4: What do you think the score represents?

E. Now I'd like you to compare actin from several organisms.

- *Homo sapiens* (human)
- *Lymantria dispar* (gypsy moth, an insect)
- *Rattus norvegicus* (Norway rat)
- *Caenorhabditis elegans* (nematode, a worm)
- *Gallus gallus* (chicken)

To find the similar sequences from a specific organism, go to the "Edit and Resubmit" link, look for the box that says "Organisms." Type in the name of the

organism of the following organism in the box and the box will auto-complete your entry.

Lymantria dispar	(Gypsy Moth)
Rattus Norvegicus	(common rat)
Caenorhabditis Elegans	(nematode worm)
Gallus Gallus	(domestic chicken)

For each species, click "Blast!" and wait for the new result to come out in the other window. Click on the first hit if there are multiple hits. Scroll down to the sequence of the protein, copy and paste the sequence to your Word file, and do not forget to give a name to each sequence.

Now you have a Word file containing five sequences. Each one of them begins with a name line. For example for Lymantria dispar actin (376 aa, AAD54427)

>L.dispar

```
      1 mcdeevaalv vdngsgmcka gfagddapra vfpsivgrpr hqgvmvgmgq kdsyvgdeaq
     61 skrgiltlky piehgivtnw ddmekiwhht fynelrvape ehpvllteap lnpkanrekm
    121 tqimfetfnt pamyvaiqtv lslyasgrtt givldsgdgv shtvpiyegy alphailrld
    181 lagrdltdyl mkiltersys ftttaereiv rdikeklcyv aldfeqemat aassssleks
    241 yelpdgqvit ignerfrcpe alfqpsflgm eangihetty nsimkcdvdi rkdlyantvl
    301 sggttmypgi adrmqkeita lapstmkiki iapperkysv wiggsilasl stfqqmwisk
    361 qeydesgpsi vhrkcf
```

III. COMPARE MULTIPLE SEQUENCES (CLUSTAL)

Now that you have collected several proteins that are all similar, we are going to use a multiple alignment program called Clustal to see how similar they are to each other overall.

A. Go to: clustalw.genome.jp/

B. You will see a big box below the instructions "Enter your sequences…" Go to your word document with all the sequences and copy them all from the first > to the last amino acid. Paste this in the box. The format we have been using is called FASTA, which is accepted by this program. Make sure you are not putting any blank lines or blanks between > and the name you give to each sequence. Also, make sure there is a blank line between each specie's sequence or else the clustalw will read it as one long sequence.

C. Click "Execute Multiple Alignment". If you see the message "The sequence length is 0" then you did something wrong.

D. After a few moments, you will see a results page that includes "alignment". The regions where all of the proteins have the same amino acids will have asterisks (*) below them. Those that do not will have colons (:) below them.

Question 5: For your alignment, in what order (top to bottom) are the sequences?

Question 6: Is this the same order they were in when you submitted them to Clustal? If not, think about why and what the sequence of alignment indicating. If your answer is yes, try to change the order of your submitted sequences, and see whether that will change the alignment, thinking about why.

E. Go to the bottom of the results page and click on the link for "Select Tree Drawing". Find the "Tree Styles" menu, select "N-J Tree," and click "Exec". This will activate a program that draws visual representations of similarity called "trees."

F. A new browser window will open with your tree in it. When you look at your tree, you will see a single "ancestor" of all of the proteins represented by a single line on the left hand side. Each branch point represents a change in a protein that leads to different subtypes. Proteins that branch apart close to the left hand side are more divergent, and proteins that branch apart close to the right-hand side are more similar to each other.

G. Your tree should contain all of your species names at the ends of branches. Print out this page and turn it in along with answers to the questions in this tutorial.