NAME:

LABORATORY 8; EXERCISE 1. DATA IN THE GENBANK DATABASE

Purpose – In this exercise, we are going to have to main goals. The first is to introduce you to the different types of DNA data that is included in the GenBank database. This will include data formats, how to "visualize" information about sequences, how we can look at different "biological features" pertaining to certain sequences and many other pieces of information we can look at related to our query sequence. And secondly, you will gain practice in searching GenBank using different types of queries. Remember, GenBank is MASSIVE and growing exponentially about every 18 months! It is critically important to be able to not only search for information, but also filter that information, especially for unwanted sequences and data.

Starting the Exercise –

1. Open your web browser and go to https://www.ncbi.nlm.nih.gov/genbank/.

   NCBI hosts a number of other bioinformatics tools and you will see tabs for many of these programs below the search bar located at the top of the page. We will be concentrating on the GenBank tab on the far left of the screen under the search bar.

   All NCBI databases can be queried through a common search interface represented by the "Search" bar at the top of the page. You can use the dropdown tab to the left of the search bar to select which database(s) you intend to search.

2. Type the following into the search bar: AB001981 and hit the "Search" button.

   From what organism does the DNA originate?

   What is the name of gene? How many base pairs are represented in this entry?

   Notice that in the heading of the page, there is information on the publication from which the DNA sequence was originally cited. This information is also linked on the right hand side of the page by clicking on the "PubMed" tab under "Related Information". Click on the PUBMED link. What is the originally cited article for this entry?

Information like the original publication is of critical importance when using GenBank. This information makes it possible to trace the source or sources of any DNA sequence investigated. As the researcher, you can now find important information regarding the experiments done to generate the sequence. You can use your judgement to decide if something is "wrong" with the sequence. Maybe this particular gene shows a strange intron/exon structure compared to closely related genes. Or maybe this sequence doesn't match ANY other known genes. This would certainly be a result worth further examination. By being able to see the original publication, it is possible to double check and verify experimental procedures. GenBank submissions grow exponentially as mentioned earlier. Sometimes mistakes are made, even honest ones like stating a gene to be of type A while in fact, it is a type B gene. You can also see more serious problems ranging from bad/wrong PCR settings, bad primers or even contamination issues that may occur during the cloning of the gene. Any of these situations can generate INCORRECT DATA. It is up to you as, as the researcher, to fully investigate your findings and take nothing at face value.

3. View and Save the Sequence Entry data in FASTA format. Under the GenBank Accession number at the top of the page, click on the "FASTA" tab.

4. At the top of the page to the right, click on the "Send to" drop down. Choose "Complete Record" and then the "Clipboard" as the destination of choice.

5. Open NotePad© (or your other favorite text-editor) and "Paste" the sequence. Do not use a Word Processor as lots of additional information such as fonts and formatting are saved along with the data. These kinds of superfluous information do not work with programs to analyze sequence data. Save the file.

<u>Now it is time to explore more about the genes that are defined in this GenBank entry.</u>

6. On the left hand side of the GenBank entry for your query, find the "CDS" tab. This is the Coding Sequence for the gene. This is the region of the gene that codes for the protein. Or, basically, the sequence you get when the exons are removed from the gene. Click on the first "CDS" listed.

   What do you observe happening in the DNA sequence of the gene?

   List the intervals provided for the protein coded for. (Hint: look beside the CDS tab for the data line that begins with join (1104..). What does this information tell you about the gene?

   Where is the start codon for this protein located? What is the stop codon?

Are introns present in this coding sequence and if so, where are they located?

7. Go back to the top of the page and click on the “Graphics” tab. This changes the view on the screen to an interactive graphical representation of the GenBank entry. The upper part of the graphic shows the entire length of the entry with bars representing the individual exons within the gene.

   You can zoom the graphic view by clicking in the upper part of the graphic and dragging the blue box over the area of interest. You can also zoom by right-clicking and using the “Zoom” feature.

   By “mousing over” the bars in the graphics, you can also find out additional information about that particular feature of the gene.

   This graphic overview of the gene is mostly useful when you are examining genes with multiple GenBank entries (sometimes this number can be in the hundreds). Play around with this interface for a few minutes to what functionality is offered through the program. You can always use the “Back” arrow to go to the preceding page or if you get lost, just reenter the original Accession number for this exercise.

8. Go back to the original page for this entry and click on the other CDS tab lower on the left side of the page. What is different about this CDS compared to the first one you looked at? Your answer to this question should be in the form of a comparison between the two based on the information you looked at for the first CDS.

9. Taking into account what your answered for question #8, what do you think this says about these two CDS?