**Bioinformatics Worksheet #2: BLAST Practice - Instructions**

**Introduction**: In this exercise, you will be assigned, based on your group number, a mystery bacterial operon in which there are multiple genes to identify. You will be using the bioinformatics tools of BLAST and ORF finder to identify those genes and explore the function of the operon.

**Procedure**:

1. Select one of the mystery bacterial operons provided by the instructor and go to NCBI’s Open Reading Frame (ORF) finder (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html> ). This program, from the National Center for Biotechnology Information (NCBI) provided by the National Institute of Health (NIH), finds possible gene sequences in a section of prokaryotic DNA. Copy and paste your sequence into the large text box and click on the **OrfFind** button.
2. **A.** Operons are transcribed from the DNA to RNA through transcription. That transcript has multiple genes, each with their own functions (though they are often interlinked). But they are often not in the same reading frame, so programs like ORF finder are used to locate possible genes. On this page, each one of the blue bars represents a possible gene sequence as, in one of the 6 frames, there is a consensus sequence close to a start codon. However, not all of these actually encode for a protein. The next step is to run the possible genes through BLAST to see if they match anything on record, or are very similar to protein belonging to a different species if a record of this protein doesn’t exist.

**B.** Start with the larger bars and work your way smaller. You can use the list on the right of the page to go in order. Generally, the actual genes aren’t smaller than 120-210 bp (encoding form a protein of about 40-70 amino acids (aa)). Click on one of the bars, or in the list to the right, then, open a new page in which to investigate your proteins by pushing and holding the **Ctrl** key on your keyboard while clicking with the left button on the mouse on the button that says **BLAST**, near the top in the pink section of the page. The window that opens is called a BLAST request (Basic Local Alignment Search Tool), which provides a pre-filled in form. To process the request, just click **View report**.

1. The BLAST search may take anywhere from a few seconds to a few minutes, so be patient. Once BLAST is done searching, identify the probable function of the protein by looking at the most similar result (These start below the box filled with red lines). Write down the function of the closest match and record the E- value, query cover, and % Identity as well as the length of your protein in both base pairs and amino acids in **Table 1**. This may not always be the first result.
2. To learn of the function of your protein, look at the other results. Often there will be a pattern in the names of the proteins in terms of function (ex. Transporter, decarboxylase, etc.). Alternatively, click on the name of your closest match. This will take you further down the page for a more detailed result. Click on the entry for sequence ID (near the top). This takes you to a page with information about the protein. Use Ctrl+F to search for the keyword function. If this doesn’t work, search for the protein by name on Pubmed (<http://www.ncbi.nlm.nih.gov/>). You should be able to figure out the function of the protein from the search results.
3. Repeat steps 2 through 4 for each of your proteins until you have found the eight largest proteins (or all of them – some operons have fewer proteins in them) in this operon and use that information to answer the following questions.

**Bioinformatics Worksheet #2: BLAST Practice - Assignment**

**Group Number:\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Names of individuals in group:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Mystery bacterial operon #\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Table 1: Table of results for 8 largest proteins in operon**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Protein #** | **Length (base pairs)** | **Length (amino acids)** | **E-value** | **Query Cover** | **% Identity** | **Protein name – brief description** |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |
| 6 |  |  |  |  |  |  |
| 7 |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |

1. What organism did this operon come from? What evidence did you use to come to this conclusion?
2. Are the functions of the genes found in this operon related? What cellular system do you think this operon and the genes contained are involved in?
3. Do some research using NCBI (<http://www.ncbi.nlm.nih.gov>) and other internet resources to learn about the system. What role does this system have in this organism? Create a hypothesis about what might happen to the organism if these genes were deactivated.
4. Share what you learned with the other groups and write down what you learned about the other systems.

**Bonus**: Go to RCSB Protein Data Bank (<http://www.rcsb.org/pdb/home/home.do>) and search for one of the proteins you found by name. You may not be able to find it for your organism, or at all, but here you can see the structure of the protein determined by protein crystallography. First, type the name of your protein (e.g. beta-D-galactosidase) in the window at the top and hit enter, then scroll down to the search results, click the first (or most relevant link), and view the **Summary** tab. Count the number of beta pleatings and alpha helixes (can also be found under the **Sequence tab**). Are there any ligands associated with this protein? Are there any other units for this protein? If so, name them. Be sure to look at the protein in the **3-D view**.

Name of protein you investigated:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Number of Beta-pleated sections:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Number of alpha helix sections:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Associated ligands?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Additional subunits for protein?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Provide a rough 3-D sketch of your protein on the back of this page: