Lab sheet #4
-Introduction to the BLAST Suite and BLASTN-

-Objective:

- To know how to use BLSTN, and find similarities between sequences.

Use NCBI database and BLASTN web page to answer the following questions:

Exercise 1: Biofilm analysis

Public water supply lines are immersed in water for decades and a community of microorganisms thrives on these wet surfaces. These slippery coatings are referred to as biofilms and the bacterial makeup is generally unknown because scientists are unable to culture and study the vast majority of these organisms in the laboratory. In 2003, Schmeisser and colleagues published a study where they collected and sequenced the DNA from bacteria growing on pipe valves of a drinking water network in Northern Germany. Through **sequence similarity**, they were able to classify a large number of these organisms as belonging to certain **species or groups**. In this process they identified many new species. In this exercise, you are asked to use **BLASTN** to repeat some of their analysis and identify the makeup of these biofilms.

Below is a list of 5 sequence accession numbers from their study. You are going to use the **NCBI BLASTN** web form to search for sequence similarities to try to identify the bacteria growing within these biofilms.

AY187314 - AY187315 - AY187316 - AY187317 - AY187318

Questions:

1. Retrieve the sequence from the **NCBI GenBank** and, based on the annotation of these sequence records, identify what gene was used in their analysis.
2. Convert the file format to **FASTA**.
3. Navigate to the **NCBI BLASTN** Web form and paste the FASTA format of each DNA sequence into the Query window. **OR** from the right menu choose RUN BLAST.
4. You can also BLAST the sequence using the accession number.
5. In BLAST:
   - Choose the "**Nucleotide collection (nr/nt)**" as the database to be searched.
   - To save lots of time for your searches, restrict your search to "**bacteria (taxid:2)**" in the **organism field**.

- Pick "**Somewhat similar sequences (blastn)**" as the program to be used in the search.
- Launch the search by clicking on the "**BLAST**" button.
- Save your **search strategy**, to BLAST the rest of the sequences.

6. Open up additional Internet browser windows and launch the other searches.

- Five individual windows of results will be returned within a few minutes. Be sure to stay organized and record your conclusions for each accession number.

7. For each BLASTN search, survey the results **graphic**, **table**, and **alignments** to assign each unknown sequence to an organism. You may not find 100% identity between your query and the hits, except for the self-hit.

- Note that the first hit may also be an unknown so you should examine all the hits before drawing any **conclusions as to what kind of bacteria the sequence came from.**

8. From the common header, brows your recent results.

Exercise 2: RuBisCO

It is often said that ribulose bisphosphate carboxylase (RuBisCO) is the most abundant protein on the planet. This enzyme is part of the Calvin cycle and is the key enzyme in the incorporation of carbon from carbon dioxide into living organisms. It is part of an enzyme complex found in plants, terrestrial or aquatic, and most probably played an important role in the development of our atmosphere and life on earth.

Arabidopsis thaliana, a member of the mustard family, is an important model system for higher plants. It is easily cultivated in the laboratory, undergoes rapid development, and produces a large number of seeds, making it amenable to genetic studies. Although not important agronomically, Arabidopsis has provided fundamental knowledge of plant biology and it was the first plant genome to be sequenced in 2000.

In this exercise, you will use **BLASTN** to identify members of the RuBisCO gene family in Arabidopsis.

Questions:

1. Run BLASTN for Arabidopsis RuBisCO small chain subunit 1b mRNA using its accession number (**NM_123204**) :
   - Set the database to "**Reference RNA sequences (refseq_rna)**" and restrict the organism to "**Arabidopsis thaliana (taxid:3702)**."
   - Set the program selection to "**Somewhat similar sequences (blastn)**" and click on the "BLAST" button to launch the search.
   - When the results are returned, you should now utilize the graphic, table, and alignments to identify the **family members.**

2. The Reference RNA database should not have any redundancy but two family members have alternatively spliced mRNAs. Compare the alignments carefully and examine the annotation (especially the coordinates of the coding regions) of all the relevant sequence records to describe and understand the major differences between these **family member transcripts**.
   - Create a table with a listing of the **names** of **family member transcripts** and their **accession numbers**, their **mRNA length**, and the coordinates of the coding regions (**CDS**).