

## Lab sheet 8

### Introduction to Sequence Alignment

#### **-Objective:**

- To know how to use BLSTN, and find similarities between nucleotide sequences.
- To know how to use BLSTP, and find similarities between protein sequences.
- To know how to use Clustal W/Omega, and find similarities between multiple sequences.

### *Nucleotide Sequence Alignment*

#### 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 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 each sequence from the **NCBI GenBank** and, based on the annotation of these sequence records, identify what gene was used in their analysis.
2. For each sequence, convert the file format to **FASTA** using the “Display Settings.”

3. Navigate to the **NCBI BLASTN** Web form and paste the FASTA format of each DNA sequence into the Query window.
  - 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.
4. 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.
5. 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**.

**Exercise 2: RuBisCO**

It is often said that ribulose biphosphate 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, 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**.
  - Compare between these family members and 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**).

## *Protein Sequence Alignment*

### **1) Protein sequence search using NCBI database and BLASTP web page:**

Navigate to the **NCBI BLASTP** web form and write the accession number (**CAG33009.1**) of homo sapiens X-ray repair cross complementing 1 (**XRCC1**) into the query window.

- Choose the “non-redundant protein sequences (nr)” as the database to be searched.
- To save lots of time, restrict your search to the organism under study in the “Organism field”. In this example we are looking for sequence similarity with **mouse** (taxid:10088).
- Pick “blastp (protein-protein BLAST)” as the program to be used in the search.
- Launch the search by clicking on the “BLAST” button.

### **2) Multiple protein sequence alignment using Clustal Omega:**

1. Navigate to the **NCBI BLASTP** web form and write the accession number (**NP\_065826.3**) of homo sapiens Estrogen-induced gene 121 protein into the query window.
  - Choose the “non-redundant protein sequences (nr)” as the database to be searched.
  - Pick “blastp (protein-protein BLAST)” as the program to be used in the search.
  - Launch the search by clicking on the “BLAST” button.
2. In BLAST result page, click **Taxonomy Reports**. Showing the similar proteins in other organisms.
3. In organism report, copy the accession number of the first hit under five interested organisms.
4. Paste these accession numbers with (**, space**) between them in NCBI protein database search box. Change the Display setting to **FASTA text** and copy all sequences.
5. Open **Clustal Omega** and paste the sequences in the input box. Change the first part of each sequence to the organism name.
6. Pick “**protein**” as the sequences to be aligned are proteins. Choose “ **ClustalW with character count**” as the output format and click **Submit**.
7. Click “**show colors**” in the result page.

8. All the sequences now are aligned under each other, with different signs represented under each amino acid position (\*, :, ., -) showing the extent of similarity.
- ✓ \* → indicates positions which have a single, fully conserved residue.
  - ✓ : → indicates conservation between groups of strongly similar properties
  - ✓ . → indicates conservation between groups of weakly similar properties
9. Download the alignment file.

### **Exercises:**

- 1- Find any sequence similarity to the human TRF2 protein (NP\_005643.2) with horses (taxid:9788).
- 2- Given are two proteins. Find out whether or not they are homologs. Test different algorithms for pairwise sequence comparisons.
- Protein 1** (*HrpB7* from *Xanthomonas campestris* pv. *vesicatoria*)
- Protein 2** (*HrpD* from *Ralstonia solanacearum*)
- 3) Construct a pairwise global alignment for **bovine chymotrypsin**, **bovine trypsin**. Based on their sequence alignment, **determine the residues for trypsin that corresponds with chymotrypsin residues 189, 190, and 228 that line the specificity pocket of the enzyme and name them.**

**Note → First make sure that the alignment is anchored properly, this is done by:**

- Determining the location of active site residues **His, Asp** and **Ser** of **bovine chymotrypsin** from Uniprot page.
- Obtaining sequences of **bovine chymotrypsin** (gi 157831162) and **trypsin** (gi 60593450) in FASTA (text) format from the **NCBI protein database**.
- Making sure that the active site **Ser, His, and Asp** of chymotrypsin is aligned with those of trypsin.