# BLAST to answer biological questions

## Introduction

*In this practical assignment you will use the BLAST programs to answer two biological questions:*

1) *Detect frameshift mutations using sequence alignment. (Task 1)*
2) *Find structurally similar proteins that have very little sequence similarity. (Task 2)*

*For each task, follow the instructions given and type your answers to the questions asked after the question.*

## Tools used in this session

*NCBI BLAST*

### Task 1: Detecting Frameshift mutations

**Task 1: instructions**

*Search the sample sequence given below against the non-redundant protein database (nr) using the BLASTX program. Leave all other input parameters as default.*

*BLASTX searches a nucleotide sequence against an amino acid database by translating the nucleotide sequence into its 6 possible reading frames.*

>sample_sequence
AGAAGAAGACATAGTAATTAGATCTGAAAATTTTACGAACAATGCTAAAACCATAATAGTACAGCTGAAG
GAATCTATAAAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAAGTATACCTATAGCAACGGGGG
GAGCAATTTATGCAACAGGAGACATAATAGGAGATATAAGACAAGCACATTGTAACCTTAGTAGAGACCA
ATGGGATAACACTTTAAGCCAGCTAGTTACAAAACTAAGAGAACAATTTGGGAATAAACAATAGCCTTT
AATCAATCCTCAGGAGGGGACCCAGAAATTGTAATGCACAGTTTTAATTGTGGAGGAGAATTTTTCTACT
GTAATACAACACAGCTGTTTAATAGTACTTGGCCAACTAATAAAAAGTCTACTAACAAAACAGGAAC
TATCACACTCCCGTGCAGAATAAAACAAATTATAAACAGGTGGCAAGAAGTAGGAAAAGCAATGTATGCC
CCTCCCATCAAGGGACAAATTAGATGTTCATCAAATATTACAGGGATATTCTTAACAAGAGATGGTGGTA
ACGCAAGCGATGAGACCGAGACCTTCAGACCTGGAGGAGGAAATA

Answer the following questions.

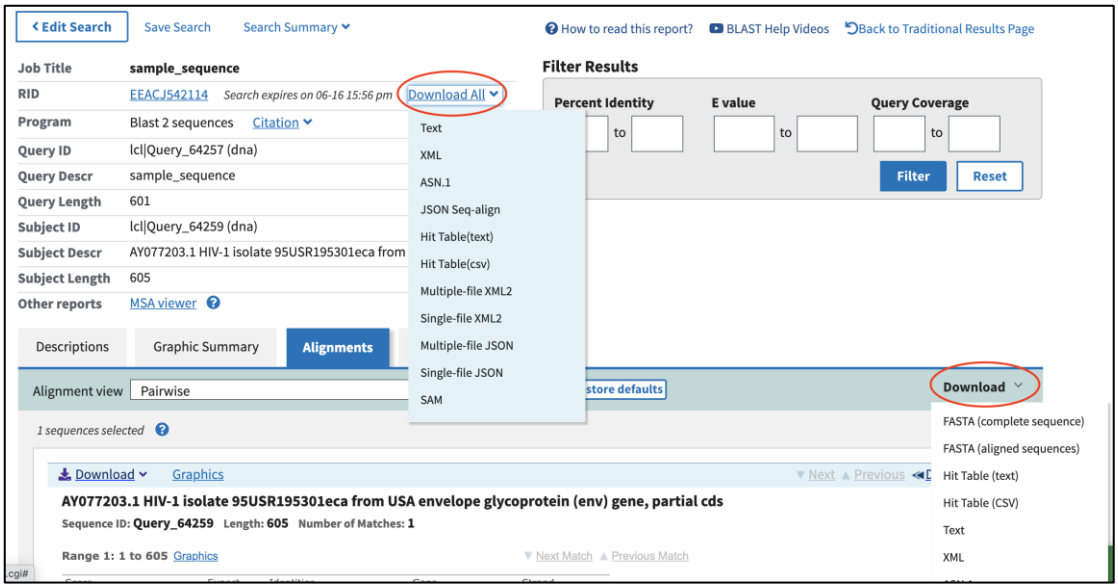
1. **For database hit AAL71600.1 in which translation frame does the query sequence align to the sequence of AAL71600.1 at the 5' end of the query sequence?** *Hint: pay attention to the alignment coordinates of the HSPs.*

2. **After which nucleotide position of the query sequence does the alignment frame change to +1?** *Hint: In BLASTX the coordinates for the query sequence are relative to*

*its nucleotide sequence position while the coordinates of the subject sequence are relative to its protein sequence position.*

3. **Using what you learnt about navigating through the NCBI database, find the nucleotide sequence corresponding to the protein with the accession AAL71600.1. Describe the steps you took to find it.**

4. **Make a local alignment of the nucleotide sequence you retrieved from 3 above with the sample sequence given above using the 'Align two or more sequences' option of BLASTN. Download your alignment and paste it below.** *Hint: To download your alignment or any search result, for that matter, click on 'Download' at the top of your results page. Select the format you would like to download your results in. Try the 'text' format for now. Alternatively, you can download your alignment results from the Alignment tab. See screenshot below. The 'Download' links are circled in red.*




**Paste your alignment below.** *Hint: to render the alignment properly use Courier New font at font size 8 or below. Or paste a screenshot.*

5. **Highlight the nucleotide that is bringing about the frameshift mutation in the alignment you have pasted in 4.** *Hint: Refer to your answer from question 2. A frameshift mutation is an insertion or a deletion (indel) that causes the protein translation frame of the sequence to shift. Indels that are not divisible by 3 usually cause a frameshift mutation.*

6. **What other differences are there between the two nucleotide sequences? Do they also cause a frameshift mutation? Do they change the amino acid i.e. are they non-synonymous changes?**

**Task 2: Finding structurally similar proteins using PSI-BLAST**

**Task 2: instructions**

*Tyrosine tRNA ligase (TyrRS) and Tryptophan tRNA ligase (TrpRS) are structurally similar (Refs: 1, 2). Given structural similarity you would expect to find sequence similarity. However, TyrRS and TrpRS share 13% sequence identity.*

*We will use PSI-BLAST to find the sequence of TrpRS using the sequence of TyrRS.*

1. **Using a sequence database of your choice retrieve the protein sequence for *E. coli* Tyrosine tRNA ligase (TyrRS), alias Tyrosyl-tRNA synthetase (sp|P0AGJ9).** *Hint: You can either use NCBI GenBank or SwissProt.*

2. **Open the BLASTP webpage on NCBI and paste the sequence you found in 1.**

3. **Submit a PSI-BLAST search against the UniProtKB/Swiss-Prot database, narrowing your search organism to Bacteria (taxid: 2). What is the default value for the PSI-BLAST threshold?** *Hint: Look at the Algorithm parameters*

4. In your search results for the first iteration, the description section will be split in two tables. The top table will contain alignments with E-values within the cut-off and the lower table will show the hits above the E-value cut-off. Answer the following questions:

    a. **What is the range of % identities for the alignments in the top table?** *Hint: sort the % identity column.*

    b. **Do you get hits to proteins other than TyrRS in the first table?**

    c. **Take a look at the Taxonomy report for this search by clicking on the 'Taxonomy' tab. Which organism has the lowest Max. alignment score for TyrRS?** *Hint: Sort the hits by Max. score in the Descriptions tab to reflect the order in the other tabs. NB: Taxonomy report contains hits from both the tables in Descriptions tab.*

5. Run the 2nd iteration of PSI-BLAST using hits in the first table. This may take a while to run so please be patient. If it takes you back to the submission page click on the browser back button and re-submit the 2nd iteration. Using the output for the 2nd iteration answer the following questions:

    a. **Are any new sequences now added to the top table? Are any of them of Tryptophan-tRNA ligase?**

**b. Save the top table in text format for future comparisons. To which organism does the best scoring Tryptophan-tRNA ligase hit belong to? What is its alignment score?**

**c. Do you get a Tryptophan-tRNA ligase hit to *Escherichia coli* in this iteration?**

6. Run a 3rd iteration with the hits in the top table.
   a. **Do you get a hit to TrpRS of *E. coli*?**
   **b. If yes, what is the alignment score and accession number of the best scoring hit?**
   **c. What % identity does the query sequence (*E. coli* TrpRS) align with it (*E. coli* TyrRS)?**