BT302: Bioinformatics Lab Assignment No. 3 24.08.2022

Note 1: Submit the assignment online through Moodle either in .doc or .pdf format. Your final report file should be named as "**YourName_BT302_Lab3_24082022**". Make sure that your name and roll numbers are written at the first page of your final report. Note that you can upload only one file; thus, put together all the answers in a single file.

Note 2: There are two parts of this assignment. In this first part, students are expected to answer the questions 1 and 2. In the second part, students are expected to write a script to achieve the intended results. Each student should choose only one of the parts.

--PART 1---

Goal of this exercise is to learn about sequence alignments using the program BLAST.

1. **Download the protein sequences of Cas9-1 and Cas9-2 of *Streptococcus thermophilus* (strain ATCC BAA-491 / LMD-9) from UniProtKB database.**

   (a) Use the program BLASTP to align the two sequences that you have downloaded.

   (b) Note down the set of parameters (e.g. e-value, word size, etc.) used to perform the pairwise alignment.

   (c) Note down the statistics of alignment (e.g. max score, total score, query coverage, e-value, identity, similarity, bit-score, gaps, etc).

   (d) Choose **different scoring matrices** and prepare a comparative table of alignment statistics. What is/are your conclusion(s)?

2. **Use the sequence given below as a query to perform the following exercises**

   ```
   >Query Sequence
   MKRNYILGLDIGITSVGYGIIDYETRDVIDAGVRLFKEANVENNEGRRSKRGARR
   LKRRRRHRIQRVKKLLFDYNLLTDHSELSGINPYEARVKGLSQKLSEEEFSAALL
   HLAKRRGVHNVNEVEEDTGNELSTKEQISRNSKALEEKYVAELQLERLKKDGEV
   RGSINRFKTSDYVKEAKQLLKVQKAYHQLDQSFIDTYIDLLETRRTYYEGPGEGS
   PFGWKDIKEWYEMLMGHCTYFPEELRSVKYAYNADLYNALNDLNNLVITRDEN
   EKLEYYEKFQIIENVFKQKKKPTLKQIAKEILVNEEDIKGYRVTSTGKPEFTNLKVY
   HDIKDITARKEIIENAELLDQIAKILTIYQSSEDIQEELTNLNSELTQEEIEQISNLKG
   YTGTHNLSLKAINLILDELWHTNDNQIAIFNRLKLVPKKVDLSQQKEIPTTLVDDFI
   LSPVVKRSFIQSIKVINAIIKKYGLPNDIIIELAREKNSKDAQKMINEMQKRNRQTN
   ERIEEIIRTTGKENAKYLIEKIKLHDMQEGKCLYSLEAIPLEDLLNNPFNYEVDHIIP
   RSVSFDNSFNNKVLVKQEENSKKGNRTPFQYLSSSDSKISYETFKKHILNLAKG
   KGRISKTKKEYLLEERDINRFSVQKDFINRNLVDTRYATRGLMNLLRSYFRVNNL
   DVKVKSINGGFTSFLRRKWKFKKERNKGYKHHAEDALIIANADFIFKEWKKLDKA
   KKVMENQMFEEKQAESMPEIETEQEYKEIFITPHQIKHIKDFKDYKYSHRVDKKP
   ```

NRELINDTLYSTRKDDKGNTLIVNNLNGLYDKDNDKLKKLINKSPEKLLMYHHDP
QTYQKLKLIMEQYGDEKNPLYKYYEETGNYLTKYSKKDNGPVIKKIKYYGNKLN
AHLDITDDYPNSRNKVVKLSLKPYRFDVYLDNGVYKFVTVKNLDVIKKENYYEVN
SKCYEEAKKLKKISNQAEFIASFYNNDLIKINGELYRVIGVNNDLLNRIEVNMIDIT
YREYLENMNDKRPPRIIKTIASKTQSIKKYSTDILGNLYEVKSKKHPQIIKKG

(a) Use the sequence to search for the homologous protein sequences in **few different databases**. Do you get different results using different databases? If yes, list some of the results which are different.

(b) Use the sequence as query and choose the UniProtKB database and perform the homology search using the program PSI-BLAST.

i. Do you get the number of hits in the first iteration same as in the previous question (2a)?

ii. Perform the second iteration and see if the hits are now different or same.

iii. If you get new hits, list them down. If you do not get new hits, give the possible reason(s).

iv. Perform few more iterations until you get no new hit. After how many iterations you got no new hit i.e. search got saturated.

v. What are the different types of organisms which contain this protein?

vi. Paste the results of distance tree in your report.

--END of PART 1-----------------------------------

--PART 2---

1. Write a code to align two user-input protein sequences globally using substitution matrices as scoring scheme.

Note: Copy paste or attach your written code in the report and create a README file on how to use the code.

--END of PART 2-----------------------------------