IN4030 - Partial exam 1

Candidate ID: 15023

Exercise 1: Database searches (8p)

a)

Retrieve the protein sequence with accession number NP_001264426 from the National Center for Biotechnology Information (NCBI) using their website at https://www.ncbi.nlm.nih.gov/ and its search functions. The sequence should be 247 amino acids long. Save it in a FASTA-formatted text file. Show the formatted sequence in the report. (1p)

```
import requests
link = ("https://www.ncbi.nlm.nih.gov/search/api/"
        "download-sequence/?db=protein&id=NP_001264426.1")
response = requests.get(link)

# The content is in binary format so we decode it with utf8
content_str = response.content.decode("UTF8")

# The data we got is already in fasta format, so we write it direclty to a fasta file
with open("NP_001264426.fasta", "w+") as f:
    f.write(content_str)
```

b)

Which variant of BLAST should be used to search the UniProKB/Swiss-Prot database with the retrieved sequence as the query? (1p)

blastp, sicne it is designed for proteins as opposed to blastn, which is designed for nucleotides

c)

Use BLAST at https://blast.ncbi.nlm.nih.gov/Blast.cgi to search the UniProKB/Swiss-Prot database with the retrieved sequence as the query. Algorithm parameters must be adjusted as follows:

```
"Expect threshold" = 1,
"Word size" = 3,
"Compositional adjustments" = "No adjustment",
"Low complexity regions" = OFF.
scoring matrix = BLOSUM62,
gap costs: Existence=11 Extension=1
```

How many hits do you get? (1p)

A: I got 10 results

d)

What is the accession number of the best match with a human sequence? Save the sequence in a file. (1p)

Q6NS38.1

e)

Show the BLAST alignment between the query protein and the best matching human protein. (1p)

```
Alignment statistics for match #1
Score   Expect  Identities  Positives   Gaps
319 bits(817)   1e-109  163/262(62%)    192/262(73%)    23/262(8%)
Query  1    MDRFVVKRS--------------AEEP---GGDG----KKPRLEEEAGGLPHPSQPS-QE  38
            MDRF+VK +               EEP   GGD     K+PR E    G  H + PS +
Sbjct  1    MDRFLVKGAQGGLLRKQEEQEPTGEEPAVLGGDKESTRKRPRREAPGNG-GHSAGPSWRH  59

Query  39   IRAQGLSLEYRLLFGRAEADAIFQQLEKEVEYFEGEQTKLHVFGKWHNIPRKQVTYGDPE  98
            IRA+GL   Y +LFG+AEAD IFQ+LEKEVEYF G   ++ VFGKWH++PRKQ TYGD
Sbjct  60   IRAEGLDCSYTVLFGKAEADEIFQELEKEVEYFTGALARVQVFGKWHSVPRKQATYGDAG  119

Query  99   LTYTYSGVTFSPKPWIPVLNHIRDRLVLETGHTFNFVLINRYKDGEDHIGEHRDDEKELV  158
            LTYT+SG+T SPKPWIPVL  IRD +   TG TFNFVLINRYKDG DHIGEHRDDE+EL
Sbjct  120  LTYTFSGLTLSPKPWIPVLERIRDHVSGVTGQTFNFVLINRYKDGCDHIGEHRDDERELA  179

Query  159  PRSPIASVSFGACRDFVFRHCDSRGKNATRHIKPIRLQLAHGSLLMMKYPTNVYWYHSLP  218
            P SPIASVSFGACRDFVFRH DSRGK+ +R +  +RL LAHGSLLMM +PTN +WYHSLP
Sbjct  180  PGSPIASVSFGACRDFVFRHKDSRGKSPSRRVAVVRLPLAHGSLLMMNHPTNTHWYHSLP  239

Query  219  IRRRVLAPRINLTFRKMMAVDK  240
            +R++VLAPR+NLTFRK++   K
Sbjct  240  VRKKVLAPRVNLTFRKILLTKK  261
```

f)

What is the score (raw alignment score), the bitscore and the E-value of this alignment? (3p)

319, 817, E-value: 1e-109

Exercise 2: Alignments (22p)

a) Write a program to compute the optimal local alignment of two amino acid sequences using the Smith-Waterman algorithm with affine gap penalties. This issue can be solved in at least two different ways. The program shall read the two sequences in FASTA format from two separate files and print the optimal alignment score and the actual alignment. You only need to show one alignment if there are multiple optimal alignments with the same score. The alignment should be presented in a comprehensible manner, preferably in a way similar to BLAST. The program shall use the BLOSUM62 scoring matrix and an affine gap penalty function with gap opening penalty of 11 and a gap extension penalty of 1, i.e. gap penalty = 11 + gap length (12p)

- You are not allowed to use libraries or other imported code for reading the FASTA files or performing the actual alignment.
- The BLOSUM62 matrix can be downloaded from the NCBI at https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt
- Ideally, the scoring matrix is read from a specified file and the gap penalties are specified as options to the program. Information about testing your program is provided on page 2

See chapter 5 in the book for details (especially lesson 12.5 "Penalizing Insertions and Deletions in Sequence Alignments" at https://www.bioinformaticsalgorithms.org/bioinformatics-chapter-5).

c) Compare thescore you got with the rawscore from BLAST in exercise1 and comment on any discrepancies. (2p)

d) Compare the actual alignment you got with the alignment from BLAST and comment on any discrepancies.(2p)

e) Retrieve the ALKB_ECOLI (P05050) sequence from Escherichia coliin FASTA format. Use your program to compare this sequence with the sequence fromexercise1a. What score do you get? Show the alignment. Why did the sequence from E. coli not showup in the search in exercise 1? (4p)