IN4030 Spring 2021 - Partial exam 1

General information

- Partial exam 1 counts 15% of the final grade. Maximum points are shown for each exercise.
- The exercises must be solved individually.
- Requirements for assignments at the Department of Informatics must be followed. The rules are here: http://www.uio.no/english/studies/admin/compulsory-activities/mn-ifi-mandatory.html
- Code may be written in any programming language you consider appropriate. Python is recommended.
- Write a short report with a description of your approach to solving each exercise, a summary of the results along the way, and the answers to all the questions. Please Courier or another monospaced font on sequences and alignments for readability.
- The source code may be embedded in the report or enclosed as a separate file (preferred).
- Remember that it is more important that your code is correct than fast.
- The report must be submitted no later than Friday 12 March 2021 at 17.00 through Inspera (https://uio.inspera.no/) in the form of one single PDF document and optionally one source code file.

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)
b) Which variant of BLAST should be used to search the UniProKB/Swiss-Prot database with the retrieved sequence as the query? (1p)
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" and Filter “Low complexity regions” = OFF. The scoring matrix should be BLOSUM62, and gap costs should be “Existence: 11 Extension: 1”. How many hits do you get? (1p)
d) What is the accession number of the best match with a human sequence? Save the sequence in a file. (1p)
e) Show the BLAST alignment between the query protein and the best matching human protein. (1p)
f) What is the score (raw alignment score), the bitscore and the E-value of this alignment? (3p)

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. 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). 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. Ideally, the scoring matrix is read from a specified file and the gap penalties are specified as options to the program. You are not allowed to use libraries or other imported code for reading the FASTA files or performing the actual alignment. Information about testing your program is provided on page 2. The BLOSUM62 matrix can be downloaded from the NCBI at https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt (12p)
b) Run the program on the query sequence and the human sequence from exercise 1. What score do you get? Show the alignment. (2p)
c) Compare the score you got with the raw score from BLAST in exercise 1 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 coli* in FASTA format. Use your program to compare this sequence with the sequence from exercise 1a. What score do you get? Show the alignment. Why did the sequence from *E. coli* not show up in the search in exercise 1? (4p)

Testing the code

Below is an example of an optimal local alignment using affine gap penalties.

The human (Homo sapiens) Ogg1 protein:

```
>OGG1_HUMAN
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQTE
EQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIEC
LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGYRARYVSA
SARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYS
WHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG
```

The Ogg1 protein from yeast (Saccharomyces cerevisiae):

```
>OGG1_YEAST
MSYKFGKLAINKSELCLANVLQAGQSFRWIWDEKLNQYSTTMKIGQQEKYSVVILRQDEENEILEFVAVG
DCGNQDALKTHLMKYFRLDVSLKHLFDNVWIPSDKAFAKLSPQGIRILAQEPWETLISFICSSNNNISRI
TRMCNSLCSNFGNLITTIDGVAYHSFPTSEELTSRATEAKLRELGFGYRAKYIIETARKLVNDKAEANIT
SDTTYLQSICKDAQYEDVREHLMSYNGVGPKVADCVCLMGLHMDGIVPVDVHVSRIAKRDYQISANKNHL
KELRTKYNALPISRKKINLELDHIRLMLFKKWGSYAGWAQGVLFSKEIGGTSGSTTTGTIKKRKWDMIKE
TEAIVTKQMKLKVELSDLHIKEAKID
```

The optimal local alignment of these sequences using the BLOSUM62 matrix, a gap open penalty of 11 and a gap extension penalty of 1 should give a raw alignment score of 462. The optimal alignment may look like this:

```
Query:  30 RSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLT-QTEEQLHCTVYRGDKSQ---- 84
           +SEL L  VL +GQSFRW       W   L     T+    +E+    + R D+
Sbjct:  12 KSELCLANVLQAGQSFRWI------WDEKLNQYSTTMKIGQQEKYSVVILRQDEENEILE 65

Query:  85 ----ASRPTPDELEA-VRKYFQLDVTLAQLYHH-WGSVDSHFQEVAQKFQGVRLLRQDPI 138
                     D L+  + KYF+LDV+L  L+ + W   D  F +++   QG+R+L Q+P
Sbjct:  66 FVAVGDCGNQDALKTHLMKYFRLDVSLKHLFDNVWIPSDKAFAKLSP--QGIRILAQEPW 123

Query: 139 ECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRK 198
           E L SFICSSNNNI+RIT M   LC  FG  +  +D V YH FP+ + L     EA LR+
Sbjct: 124 ETLISFICSSNNNISRITRMCNSLCSNFGNLITTIDGVAYHSFPTSEELTSRATEAKLRE 183

Query: 199 LGLGYRARYVSASARAILEEQG------GLAWLQQL-RESSYEEAHKALCILPGVGTKVA 251
           LG GYRA+Y+  +AR ++ ++           +LQ + +++ YE+  + L    GVG KVA
Sbjct: 184 LGFGYRAKYIIETARKLVNDKAEANITSDTTYLQSICKDAQYEDVREHLMSYNGVGPKVA 243

Query: 252 DCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKG----------PSPQTNKELGN 301
           DC+CLM L     VPVDVH+  IA+RDY      +  K               + N EL +
Sbjct: 244 DCVCLMGLHMDGIVPVDVHVSRIAKRDYQISANKNHLKELRTKYNALPISRKKINLELDH 303

Query: 302 FFRSL---WGPYAGWAQAVLFSADL 323
                L   WG YAGWAQ VLFS ++
Sbjct: 304 IRLMLFKKWGSYAGWAQGVLFSKEI 328
```

Remember that there are often several optimal alignments with the same score. Also, if your program obtains the correct score and a correct alignment based on this example, it does not prove that it is correct in all cases.

It is highly recommended to check that the resulting alignment actually obtains the score that was calculated initially.

You could also test your program by comparing the results to another Smith-Waterman implementation, like the EMBOSS Water service that can be found at http://www.ebi.ac.uk/Tools/psa/emboss_water/ provided by EBI. However, with this service it is impossible to select all combination of gap penalties, so you cannot select 11 and 1 as the gap open and extension penalties, but you can compare your results to EMBOSS Water using other gap penalty settings (e.g. 10 and 1). It is important to note that EMBOSS Water includes the penalty for the first gap symbol in the gap open penalty, so an EMBOSS Water gap open penalty of 10 corresponds to a gap open penalty of 9 in BLAST and in your program.