INF4350 Partial exam no 1

General information

- This partial exam has to be accomplished individually.
- The enclosed statement "Krav til innleverte oppgaver ved Institutt for informatikk (Ifi)" must be read and accepted.
- The programs may be written in any programming language you find appropriate.
- Write a short report with a description of your approach to solve the problems, a summary of results gathered during the work, answers to the questions, as well as the all source code for the programs written.
- The report must be delivered on Wednesday 15 October 2008 at 2359 at the latest, by email to torognes@ifi.uio.no in the form of a single document file (preferably PDF).
- Direct any questions about the problems to the above email address.

Problem 1: Gene finding

a) Write a program that reads a long DNA sequence in FASTA format from a file and finds all open reading frames that may code for proteins of at least 100 amino acids. Ignore shorter reading frames. Use the universal genetic code. The program should output the following information for each of the reading frames: start position, stop position, direction, length and the corresponding translated amino acid sequence. It must also output the total number of reading frames found.
b) Retrieve the entire genome sequence of the bacterium *Helicobacter pylori* J99 from GenBank. The sequence has accession number AE001439. Download it in FASTA format. Run your program on the downloaded sequence. Show an excerpt of the results (for instance information about the first 20 reading frames found).
c) How many reading frames did you find in total?
d) How many proteins are known from this bacterium? Compare this with the number of reading frames that the program found and explain the reason for any differences.

Problem 2: Database searches

a) One of the proteins of *H. pylori* that you perhaps found solving the previous problem exists in NCBI databases with accession number NP_223766. Retrieve this sequence and store it in a FASTA formatted file. Show this sequence.
b) Use NCBI BLAST on the net at http://blast.ncbi.nlm.nih.gov/Blast.cgi to search with this protein as query against the SWISSPROT database. Under "Algorithm parameters" you should turn off "Compositional adjustment" of the score matrix ("No adjustment") and turn on filtering of "Low complexity regions". Otherwise, keep all the default parameters, specifically the BLOSUM62 score matrix, the gap open penalty of 11 and the gap extension penalty of 1. Show a brief excerpt of the results of the search.

c) How many statistically significant matches did you find?
d) Based on the results of the BLAST search, describe briefly the most probable biological function of the protein.
e) You should have obtained at least one significant match against a human sequence. What is the symbol and name of this human gene? Store the sequence of the best human match in a FASTA format in a file.
f) Show the alignment of the bacterial protein and the human protein obtained by BLAST.
g) What is the score (raw alignment score, not normalized bit score) and E-value of this alignment?
h) On which human chromosome is the gene that codes for this protein located?

Problem 3: Alignments

a) Write a program to calculate the optimal local alignment score for two amino acid sequences using the Smith-Waterman algorithm. The program should read the two sequences in FASTA format and output the optimal alignment score. The program must use the same scoring parameters as BLAST (score matrix BLOSUM62, gap open penalty 11, gap extension penalty 1). It is not necessary to store or show the actual alignment, just the score.
b) Run your program on the bacterial and human protein sequences from the previous problem. What is the score of the alignment calculated by your program?
c) Explain the reason for any differences from the score obtained by BLAST in the previous problem.

Institutt for informatikk

Krav til innleverte oppgaver ved Institutt for informatikk

Ved alle pålagte innleveringer av oppgaver ved Ifi – enten det dreier seg om obligatoriske oppgaver, hjemmeeksamen eller annet – forventes det at arbeidet er et resultat av studentens egen innsats. Å utgi andres arbeid for sitt eget er uetisk og kan medføre sterke reaksjoner fra Ifis side.

Derfor gjelder følgende:

1. Hvis du tar med tekst, programkode, illustrasjoner og annet som andre har laget, må du tydelig merke det og angi hvor det kommer fra.

2. Det er greit å få hint om hvorledes en oppgave kan løses, men dette skal eventuelt brukes som grunnlag for egen løsning og ikke kopieres uendret inn.

3. Kursledelsen kan innkalle studenter til samtale om deres innlevering.

Gruppearbeid

I noen kurs skal det leveres gruppearbeid. Ifi krever da at alle medlemmer av gruppen kan gjøre rede for hovedtrekkene i det innleverte arbeidet. Dessuten må alle ha utført en rimelig del av det hele, og kunne identifisere og svare i detalj for sin del.

Samarbeid

Reglene om kopiering betyr ikke at Ifi fraråder samarbeid – tvert imot, Ifi oppfordrer studentene til å utveksle faglige erfaringer om det meste. Men det kreves som nevnt at man kan stå inne for det som leveres.
Hvis du er i tvil om hva som er lovlig samarbeid, kan du kontakte gruppelærer eller faglærer.

www.ifi.uio.no/studinf/skjemaer/erklaring.pdf

27. jan. 2004