Due Date: Thursday, March 21, 2019 (23:55)

CENG 465
Introduction to Bioinformatics
Spring 2018-2019

Assignment #1

Programming Assignment on Sequence Alignment

Aligning RNA-Seq short reads to a reference gene

Given a set of short reads from an RNA-Seq experiment and the reference sequence of a gene, **G**, your goal in this assignment is to write an aligner to find out how much gene **G** is expressed and locate the locations of its exons. You will use semi-global alignment by dynamic programming, i.e., the Needleman-Wunsch algorithm in which the terminal gaps are not penalized, to align each short read to the reference gene. If a short read aligns to a region of the gene with a good score, you will increase the read counts for that region. After aligning each of the short reads, an integer count will have accumulated at each base of the reference gene. You will analyze this "read coverage" and write a short report about whether the gene is expressed and try to guess its number of exons.

Specifically, below are the expected step by step tasks from in this submission:

1. Get the set of short reads as a fasta file from:
   http://user.ceng.metu.edu.tr/~tcan/ceng465_s1819/shortReads_small.txt
2. Get the reference sequence of the gene that we are interested in from:
   http://user.ceng.metu.edu.tr/~tcan/ceng465_s1819/gene_hg19_refseq.txt
3. The short reads file contains about 2.4 million short reads (each read around 30 nucleotides). The reference sequene of the gene is the sequence of the gene on the chromosome from 5' Untranslated Region (UTR) to the 3' UTR and includes intronic regions. The length of the reference gene sequence is 4197 nucleotides. You will align each of the 2.4 million reads one by one to the reference gene sequence using semi-global alignment. You will use **a match score of +1, a mismatch penalty of -3, and a linear gap penalty of -2** in your alignment. Terminal gaps will not be penalized. Short read sequences contain the letter 'N' in addition to the A, C, G, and T nucleotides. **Matches to the letter 'N' will not be rewarded or penalized, they will be scored as 0 (zero)**.
4. When a short read aligns to the reference gene with a **score of +30 or more**, you will consider the short read as a mapped read and increase the coverage amount (which is the count of mapped reads) of the aligned region on the reference gene. In particular, you will maintain an array of 4197 integers and each position will indicate the number of mapped reads to that position. For example, if a short read aligns to the nucleotides 1360 to 1395 of the reference gene with a total score of 32, you will add the +1 to the coverage array from positions 1360 to 1395. Your program will output the final coverage array of 4197 integers as output (1 integer per line).
5. Analyze the coverage array by generating a line graph of the 4197 integers and discuss whether this gene is expressed and comment on its exonic regions in a 1 page short report.

You may write your code in any programming language of your choice.

Submission

Submit your source code and short report as a single .zip file via ODTU-Class before the deadline. Late submission is -20 pts per day.