Programming Assignment 4 - An exercise in isoform quantification
Computer Science CM122

Due: Friday, June 7th, 11:59pm

Introduction

This is the first and shortest of three RNA Sequencing assignments. The assignments share a common genome of 100 million bases. The three assignments also share a common set of reads. This assignment was designed to help you understand the basics of *isoform quantification*, which is critical for deciphering gene function and regulation. As you work on this small assignment, think of how you would scale up the process for hundreds of genes.

The files you need for to complete this assignments:

1. Genome file - *full_genome.txt*
2. File containing single-ended reads of length 50 base pairs - *shuffled_reads.txt*
3. Gene annotation file - *DATA_PA_1000_0*

File Formats

- *full_genome.txt* file has the the entire genome in one string.
- *shuffled_reads.txt* file contains one read of length 50 bases in each line.
- *DATA_PA_1000_0* contains the following:
    - N -- the first line in the file, indicating the number of genes in the file
    - Then for each gene i:
        - e_i -- number of exons for gene i
        - s_i_1 s_i_2 ... s_i_M -- the starting index of each of the exons
        - e_i_1 e_i_2 ... e_i_M -- the ending index of each of the exons
        - l_i -- number of isoforms for gene i
        - then l_i lines containing combinations that show which exons are in each isoform, e.g.:
            - 2 3
            - 1 2
    - Note that the genome is indexed starting at 0 and the coordinates of the exons are inclusive. E.g: An exon with start coordinate 4 and end coordinate 10 contains bases 4, 5, 6, 7, 8, 9 and 10.

These are examples that might appear as the first two genes in the file *DATA_PA_1000_0.txt*:

5
3
20377232 20379112 20384140

20377571 20379487 20384531
1
0 1 2
5
34667819 34671754 34674966 34678590 34680164
34668189 34672086 34675227 34678971 34680337
1
1 3

From this data, we see that N=5, and for the first gene i=1, we have e_1 = 3, s_1_1 = 2037232, s_1_2 = 20379112, s_1_3 = 20384140, e_1_1 = 20377571, e_1_2 = 20379487, e_1_3 = 20384531, l_1 = 1, and exons 0,1,2 appear in isoform 1. For the second gene i=2, we have e_2 = 5, s_1_1 = 34667819, and so forth.

For quantification, you need to align the reads to the exon regions listed in *DATA_PA_1100_0.txt* and count the number of reads that map onto each exon. Then use method of least squares in the procedure outlined in class to solve for the isoform frequencies.

Output File Format

The first two lines of your file should be:

>hw4_r_4_chr_1
>RNA

The solution should then be formatted with the transcript sequence and abundance in each line separated by one space. The following is an example:

>hw4_r_4_chr_1
>RNA
agcttcaaaa….. .7
gggtcaattttg…. .3
cattggaaac…. .1
tttggaccaac… .1
ggggggggcct... .8

Grading

The score will be based on the accuracy of the transcript sequence and the abundance estimate. For full credit, you should be able to get 100 both for the transcript accuracy and the abundance. This should be relatively easy to do for this project. Your grad will be the average of the two scores. Maximum grade is 100 and lowest is 0.