Programming Assignment 5 - Isoform quantification with missing information
Computer Science CM122

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

Introduction

This assignment is an extension of Assignment 4, and consists of two parts.

- **Part 1**: The first part of the assignment is to scale up what was done with 5 genes in the previous assignment to what is now 455 genes. The genes are annotated in the same format as Assignment 4. You can start by counting the reads that align to the exons and junctions to quantify the isoforms.
- **Part 2**: The goal of the second part of the assignment is to quantify the isoform abundances as well. However, the annotation will not be complete. Each gene will have an unknown exon, and an unknown isoform which contains the unknown exon. You can map the reads back to the genome to find the missing exons and quantify the isoforms as you did in Part 1 of this assignment.

File Formats

To complete the second RNA Seq assignment, you need four files.

- The *full_genome.txt* file and the *shuffled_reads.txt* file are the same as the previous assignment so you can use the same files
- There are two additional annotation files specific to this project, which are described below:
- *DATA_PA_5_1100_1* includes 455 genes. The data has the same structure as the data in the first RNA Seq assignment:
    - 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
- *DATA_PA1100_2* includes 500 genes, but not all the exons are known, and an isoform containing the unknown exon is also excluded from the data. The format is:
    - N - the first line of the file indicates the number of genes in the file

- Then for each gene i:
  - e_i -- number of known exons for gene i
  - s_i_1 s_i_2 ... s_i_M -- the starts of each of the known exons
  - e_i_1 e_i_2 ... e_i_M -- the ends of each of the known exons
  - l_i -- number of known isoforms for gene i
  - then l_i lines containing combinations that show which of the known exons are in each known isoform, e.g.:
    - 2 3
    - 1 2
  - Note that the exon which is not annotated in the data is only part of the one isoform that is not annotated. All the annotated isoforms are fully specified.

Output File Format

Your goal is to find abundance estimates for each transcript. Output should be in the same form specified in Assignment 4. You need an output file for both Part 1 and Part 2 of this assignment (i.e., submit two files). The output file for Part 1 should start with '>hw5_r_5_chr_1' and the output file for Part 2 should start with '>hw5_r_5_chr_2'. Both output files should start

The solution should be formatted with the transcript sequence and abundance in each line separated by one space. Please make sure that the special character '\n' is inserted after the abundance estimate to indicate a new line. The following is an example for Part 1:

>hw5_r_5_chr_1
>RNA
agcttcaaaa….. .7
gggtcaattttg…. .3
cattggaaac…. .1
tttggaccaac… .1
ggggggcct... .8

The formatting for Part 2 would be exactly the same but the first line should instead be
>hw5_r_5_chr_2

Grading

Part 1 and Part 2 will each count for 50% of the total grade. The score will be based on the accuracy of the transcript sequence and the abundance estimate in both parts. For full credit, you should be able to get 100 both for the transcript accuracy and at least 70 for the abundance in Part 1 and at least 70 for both transcript accuracy and abundance in Part 2.