BIOSC 1540 - Computational Biology

Transcriptomics Quiz

Mar 7, 2024

50 points

Please read the following instructions carefully before beginning your assessment.

- **Time limit:** You have 75 minutes to complete and turn in this assessment.
- **Closed notes:** You may not use additional materials for this assessment.
- **No digital devices:** No digital devices are allowed during the assessment.

I agree to the above instructions. I affirm that I will not give or receive any unauthorized help on this assessment and that all work will be my own. Your assessment will only be graded if you write your name, sign, and put your student ID below.

_____ _____

Name Signature

_____

Student ID

Problem 1

You are analyzing bulk RNA-seq data from a study comparing gene expression in diseased versus healthy tissue samples. You notice that a specific gene has a high fold change (i.e., it is up-regulated), but the large p-value indicates that it is insignificant. What could be the most likely reason for this observation? (2 points)

(A) The gene under investigation has a bimodal expression pattern across different individuals.

(B) The gene expression varies significantly within the sample groups.

(C) The sequencing depth was insufficient.

(D) The normalization method used for the gene expression data was not appropriate.

Problem 2

When comparing RNA-seq data from two different developmental stages of an organism, you find many genes with altered expression. Which factor should be considered before attributing these changes to developmental processes? (3 points)

(A) Batch effects or variations in sequencing depth between the samples.

(B) Exclusive reliance on fold-change values.

(C) Attributing all changes in gene expression to transcriptional regulation.

(D) All of the above.

Problem 3

Given the challenge of genomic variability and sequencing errors in read mapping, which approach is most effective in distinguishing true splice junctions from artifacts? (3 points)

(A) Disregarding all reads with mismatches or gaps as likely errors.

(B) Applying a uniform threshold of sequencing quality scores across all reads.

(C) Employing statistical models that account for sequencing error rates and genomic variability.

(D) Limiting the analysis to regions of the genome with little known variability.

Problem 4

The computational inference of splice junctions from RNA-Seq data involves aligning short reads that may span exon-exon junctions. This process is complicated by the vast diversity of potential splicing events and the need for high accuracy in distinguishing actual splice junctions from sequencing errors or genomic variations. Given these challenges, which strategy is most effective for improving the accuracy of splice junction identification? (2 points)

(A) Use only known splice junctions from curated databases to guide the alignment process.

(B) Use a hybrid approach that combines alignment to a reference genome with de novo assembly of reads.

(C) Increasing the stringency of alignment parameters to exclude all reads that do not perfectly match the reference genome.

(D) Utilizing a uniform threshold for read alignment across all genomic regions.

Problem 5

Aligning short reads to a reference genome presents significant challenges, especially in the context of repetitive sequences or highly variable regions. Which approach offers the best potential to enhance read mapping accuracy in these complex genomic landscapes? (3 points)

(A) Focusing only on the longest reads available in the dataset to improve the probability of unique alignment.

(B) Initially aligns reads to a simplified model of the genome and gradually integrate more complex regions.

(C) Switching to physical mapping techniques, such as optical mapping, for alignment.

(D) Enhancing read alignment with cross-species genomic comparisons to identify conserved regions.

Problem 6

What property of the Burrows-Wheeler transform is most crucial for improving the efficiency of pattern matching in biological sequences? (3 points)

(A) Its ability to compress data without losing information.

(B) The transformation of the sequence into a sorted array.

(C) The rearrangement of characters to bring similar characters together.

(D) Its reversible transformation property without needing to store the original sequence.

Problem 7

Why is the Burrows-Wheeler transform significant for bioinformatics applications in the context of FM-indexes? (3 points)

(A) It allows for efficient backward search, reducing the time complexity of finding patterns.

(B) BWT-based indexes require less storage space, making them ideal for large genomic databases.

(C) It simplifies the DNA sequence, making it easier to interpret biologically significant patterns.

(D) Both A and C are correct.

Problem 8

When applying the Burrows-Wheeler transform to a sequence, what is the importance of the last column? (3 points)

(A) It is for reconstructing the original sequence.

(B) It is important for initial pattern search steps.

(C) The property that the last column will contain all characters of the sequence.

(D) It indicates the end of a sequence.

Problem 9

Which of the following best describes the role of the Burrows-Wheeler transform (BWT) in the FM index? (3 points)

(A) It is used after the FM index is constructed to optimize search queries.

(B) BWT is the first step in FM-index construction.

(C) The transform is used for data compression before indexing.

(D) BWT provides a transformed sequence to perform rapid pattern matching.

Problem 10

The true transcriptome of a sample is defined as: (4 points)

(A) The set of all proteins encoded by the genome.

(B) The complete set of RNA molecules, including all isoforms present in the sample.

(C) The subset of RNA molecules that are being actively translated.

(D) The total DNA content of a cell.

Problem 11

The concept of effective length in RNA-seq data analysis accounts for: (3 points)

(A) The actual physical length of the RNA molecule.

(B) The theoretical maximum length of any RNA molecule.

(C) The adjustment for the empirical distribution of fragment lengths obtained during sequencing.

(D) The number of exons in the RNA molecule.

Problem 12

What is the primary goal of using maximum likelihood estimation in Salmon for RNA-Seq data analysis? (3 points)

(A) To maximize the computational efficiency required for data analysis.

(B) To maximize the accuracy of identifying RNA sequences.

(C) To maximize the probability of the observed RNA sequencing data.

(D) To maximize the amount of RNA sequencing samples.

Problem 13

Given the string “GACTTC”, determine the Burrows-Wheeler transform matrix and also indicate the resulting Burrows-Wheeler transform string. Ensure to show your work. (8 points)

Problem 14

Given the final Burrows-Wheeler transform string “G$CCGGTC”, what is the original string? (7 points)