BIOSC 1540 - Computational Biology

Quiz 03

Mar 18, 2025

20 points

Please read the following instructions carefully before beginning your assessment.

- **Time limit:** You have 15 minutes to complete and turn in this assessment.
- **Closed note:** You may not use any notes or additional resources during this assessment.
- **No digital devices:** The use of digital devices, including calculators, is not allowed.

I agree to follow the above instructions. I affirm that all work on this assessment will be my own and that I will not give or receive any unauthorized assistance. To have your assessment graded, you must write your name, sign, and provide your student ID below.

KEY _____

Name

KEY _____

Signature

KEY _____

Student ID

Problem 1

Given the reference genome of `ATCGATCGA`, construct a k-mer index using a hash map and k of 3.
(1 point)

```
{"ATC": [0, 4], "TCG": [1, 5], "CGA": [2, 6], "GAT": [3]}
```

Problem 2

How does the value of k in k-mer hashing influence read mapping results? Provide a scenario illustrating the effects of using very short (e.g., 3) or very long (e.g., 50) k-mer lengths.
(1 point)

Choosing a small k (e.g., $k = 3$) increases sensitivity by matching more locations in the genome but reduces specificity, causing many false positives due to common k-mers. Conversely, choosing a large k (e.g., $k = 50$) increases specificity, reducing false positives but risks missing true matches due to sequencing errors or genomic variations. For instance, with short k-mers like `ATG`, the genome may have hundreds of matches; with longer k-mers like `ATGCGTATCGGATCGTAGCT`, matches become rare, but small sequencing errors can prevent matching altogether.

Problem 3

Given the original string `BIOINFO`, demonstrate how to generate the Burrows-Wheeler Transform (BWT) step-by-step. Clearly show all steps and circle the final transformed BWT string.
(2 points)

Step 1: Cyclically rotate

```
BIOINFO$
IOINFO$B
OINFO$BI
INFO$BIO
NFO$BIOI
FO$BIOIN
O$BIOINF
$BIOINFO
```

Step 2: Sort lexicographically

```
$BIOINFO
BIOINFO$
FO$BIOIN
INFO$BIO
IOINFO$B
NFO$BIOI
O$BIOINF
OINFO$BI
```

Step 3: Take the last column

```
O$NOBIFI
```

Problem 4

Pseudoalignment quickly identifies which **TRANSCRIPT** a sequencing read may originate from without determining the read's exact **POSITION** within the sequence.
(2 points)

Problem 5

Which of the following best represents the role of generative models in RNA quantification?
(3 points)

(A) To detect novel transcripts from RNA-seq data.

(B) To pseudoalign sequencing reads to genomic positions.

(C) To model sequencing errors from RNA-seq data.

(D) To statistically explain how RNA fragments are sampled.

Problem 6

Which statement best explains why Salmon accounts for positional bias?
(2 points)

(A) Fragments from transcript edges are inherently less abundant biologically.

(B) Fragments from transcript edges are less likely to be sequenced.

(C) Edge fragments often contain sequencing errors.

(D) Transcript edges are often GC-rich.

Problem 7

Select all statements correctly describing the transcript-fragment assignment matrix (Z):
(1 point)

(A) Z is only used once before quantification.

(B) Z helps estimate transcript abundances.

(C) Z explicitly encodes sequencing errors.

(D) Z encodes summarizing read compatibility.

Problem 8

What is the differential gene expression analysis null hypothesis (H_0)?
(3 points)

(A) The gene is expressed equally between conditions.

(B) The gene is expressed more in one condition than the other.

(C) The gene expression is significantly different between conditions.

(D) The gene is not expressed in either condition.

Problem 9

Which of the following is the best way to increase statistical power (i.e., confidence) when performing differential expression analysis?

(3 points)

(A) More biological replicates.

(B) Increasing sequencing depth.

(C) Perform more hypothesis tests.

(D) Use normalized count data.

Problem 10

The Poisson distribution assumes that the _____**VARIANCE**_____of the data is equal to the **MEAN**_____.

(1 point)

Problem 11

What role does the dispersion parameter (α) play in the Negative Binomial distribution?

(1 point)

The dispersion parameter (α) controls the extent to which the variance exceeds the mean. Higher values of α indicate greater variability, allowing the model to accommodate overdispersion in RNA-seq count data.

Doodle area or **puzzle**: You are outside a room with a single light bulb inside, and three switches are in front of you. One of the switches controls the light bulb, while the other two do nothing. You can flip the switches as much as you want, but you may only enter the room once. How do you determine which switch controls the light bulb?

Turn on the first switch and leave it on for about 5 minutes. After 5 minutes, turn off the first switch and immediately turn on the second switch and quickly enter the room.

- If the light bulb is on, then the second switch controls it.
- If the light bulb is off but warm, then the first switch controls it.
- If the light bulb is off and cold, then the third switch is the one that controls it.