Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

PLPTH813 Bioinformatics Applications

Midterm exam

From 10:30am to 1pm

3/11/2021

This is an open book exam and the exam will take 2 hours.

**Problem 1, (10 points)**

1) Describe what these regular expressions mean.

1. [WILDCAT]{2,10}
2. 20[01][0-9]
3. pair[12].\*\.fastq
4. ^[0-9]+$
5. ^\$

**Problem 2, (15 points)**

1) Write a command to extract a second column of a file and determine the number of lines on this column containing the word “crazy” by using “cut” and “grep”. Assuming the input file, named “in.txt”, is a tab-separated text file. (3 points)

2) Write a command to concatenate two files, f1.txt and f2.txt; also write another command to merge each line in f1.txt and f2.txt. (4 points)

e.g., f1.txt: a a f2.txt: c c

b b d d

1) concatenate two files:

a a

b b

c c

d d

2) merge lines:

a a c c

b b d d

3) The full path of current directory is “/home/me/food/desserts”. Write down commands for each of the following tasks: (8 points)

1. go to the directory of /home/me/
2. check the file content in the directory of /home/me/food/
3. create a new directory “soup” under /home/me/food/
4. copy a file “copyme.txt” from the directory “/home/me/food/desserts” to “/home/me/food/soup”
5. delete the directory of “soup” and all the files in this directory

**Problem 3, (15 points)**

Two vectors were defined using the following code. The flower vector includes five flower names. The price vector includes the flower cost for each in the flower vector.

flower <- c("rose", "tulip", "lily", "Daisy1", "daisy2")

price <- c(10.00, 8.00, 12.50, 2.00, 5.50)

Write down R commands for the following jobs:

1. Identify which elements in the flower vector containing "daisy" without considering the case type (that means to identify both Daisy and daisy).
2. Change "Daisy1" to "daisy1" and assign the new vector to "flower2"
3. Determine the average price of all the flowers
4. Create a data frame containing two columns, one is for the vector “flower2” and the other is for the vector “price”
5. Determine the average price for two daisy flowers using R commands

**Problem 4, (10 points)**

Below is a DNA sequence with PHRED quality scores:

G C A C G T A T G G

38 36 35 36 28 18 14 16 17 6

Apply a window scan method to perform quality trimming using the following parameters:

Window size = 4

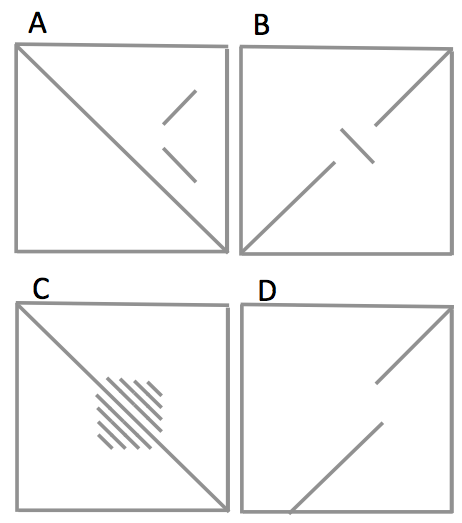
Step size = 1

Minimum quality = 20

What does the quality score of 20 mean in term of sequencing error probability? Describe the procedure of quality trimming and determine the low-quality sequence region to be trimmed.

**Problem 5, (15 points)**

Explain what you observe for each of the following dotplots of the comparison between two DNA sequences (sequence x on the x-axis and sequence y on the y-axis) **(10 points).**



What the pattern on the figure A you observed implies? (**5 points**)

**Problem 6, local alignment (10 points)**

Apply Smith-Waterman algorithm to find the best-scored local alignment between two sequences: CGCACTCGT

CGCATCGC. Here is the score scheme:

Match +1

Mismatch -1

Gap -2

Fill in score in each cell in the table and draw the path of the best-scored local alignment.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | C | G | C | A | C | T | C |
|  |  |  |  |  |  |  |  |  |
| C |  |  |  |  |  |  |  |  |
| G |  |  |  |  |  |  |  |  |
| C |  |  |  |  |  |  |  |  |
| A |  |  |  |  |  |  |  |  |
| T |  |  |  |  |  |  |  |  |
| C |  |  |  |  |  |  |  |  |
| G |  |  |  |  |  |  |  |  |

Report the best local alignment(s):

e.g. ACTGG

|||

CTG

**Problem 7, (10 points)**

Below is an alignment output using BWA:

Read10/1 99 chr1 3809 60 100M = 3973 265 GCTCCTTTATAGTATAAA

TGTAGCATATAAATATATTCCACATAAAACCATATTTGATTAATTGATCTTTGTCTAAATTACAATTATTAGAATGGAATTC CCCFFFFFHHHHHGIIGIJIJHIJIIJJJJJJJJJJIJIIJI

IJJIIJJIIIJIJJJIIJJ=EFGIIIIIJJJJGIGEEHGCACHIIGIIHHHG<CBEED NM:i:3

Read10/2 147 chr1 3973 60 28M1D72M = 3809 -265 ACTAAAACAT

GGACAAGTAGTAAAAGCTAAAAAAGTCAGAACAACTTTAGTTTAACAGTAGCTTTTGCTACCTACCTAGAATCCTAGATATACAAATCAT DDDDEEEEEEEFFFFFFHGGHHGHIIIJIJIJJJ

JJIJJIGIIIIJJJJIJJJJJJJJIGJJJJJJIHFCJHHHIJJJJJJGFJJJJHHHHHFFFFFCCC NM:i:1

Refer to the lecture notes and **interpret** the alignment result based on the first 9 columns of the alignments of this pair of reads using BWA. Please provide the detail to explain what 99 and 147 represent in the flag column (2nd column); what the information the 6th column (e.g., 100M) and the 12th column (e.g., NM:i:3) provides.

**Problem 8, (10 points)**

Describe your understanding about the sequencing technology platforms: Illumina and PacBio with respect to read length, sequencing quality (error rate and error type), and throughput. If you are interested in sequencing an unknown plant you find on campus and comparing genomic data with the public genome database to identify the species with the closest genome. Think about what technology will you choose to sequence.

**Problem 9 (NGS and alignment), (5 points)**

What software packages and what commands will you use if you would like to perform the conversion between FASTA and FASTQ? What information will you lose when you convert FASTQ to FASTA? What commands for the conversion between SAM and BAM and what information you will lose when you convert SAM to BAM?