Alexandria University
Faculty of Engineering

Bioinformatics
Spring 2024
Lab 3
Assigned: 2/3/2024
Due: 9/3/2024

Lab 3 (100 points)

Objective

Introduce students to preprocessing and sequence alignment techniques using the R programming language in bioinformatics.

Installation

To begin this lab, you need first to install R binaries and R Studio. You can find the installation instructions in the following link

Requirements

- You are obligated to attempt all tasks
- You **MUST** use R Markdown to submit your report
- State your name and ID at the first cell in your markdown report
- The lab should be

Note

For the easiness and readability of your code. Make sure to include and run this cell at the beginning of your markdown

```{r, message=F}
 # install.packages("rentrez")
 # install.packages("seqinr")
 library(tidyverse)
 library(rentrez)
 library(seqinr)
 library(Biostrings)
 library(ggplot2)
```

Part 1: Principal Component Analysis (PCA) (30 points)

For this task, use the minified version of brain cancer dataset from this link

Task 1.1: Perform PCA (30 points)

1. Perform PCA using the princomp function
2. Calculate the variation explained by each principal component
3. Plot 3 scatter plots. (Comp.1 vs Comp.2, Comp.1 vs Comp.3, Comp.2 vs Comp.3). Which plot do you think is best?
4. Draw a scree plot using the ggplot2 for the first 20 principal components

Part 2: Statistical Testing (25)

For this task, use the diabetes prediction dataset from this link

Task 2.1: Fisher's Test (15)

1. Count the alleles with respect to diabetes
   (hint 1: Use tables)
   (hint 2: Split the alleles into two columns and create two counts with each column then add them)
2. Run fisher test and report the p-value. Is the treatment significant?

Task 2.2: T Test (10)

1. Extract two BMI samples from two of the alleles family (AA, AC, CC)
2. Perform T-Test and report the p-value. What is this test measuring and what is its significance?

Part 3: Sequence Alignment (40)

Task 3.1: BLAST (10)

1. Visit BLAST from this link
2. Use the following sequences for your experiment

SEQ.A
GGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCT
TCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACA

SEQ.B
GGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACACTTGCT
TCTGACACAACTGTGTTCACGAGCAACCTCAAACAGACA

3. Report the algorithm parameters, score, percent identity, and e-value
4. How many base pairs are mismatched?

Task 3.2: Retrieve Sequences (10)

1. Install and load the necessary packages: “BioStrings” and “rentrez”

2. Fetch two sequences from GenBank (using accession numbers NG_050578.1 and X03562.1) :
   Hint: The function you'll use from the rentrez package has parameters for specifying the database, the ID of the sequence, and the format you want the sequence in. Think about how you specify the database for nucleotide sequences and the format for FASTA.

3. Store the sequence in an appropriate R data structure:
   Hint: When you fetch data, it might not be in a format directly readable by the Biostrings functions. Consider how you can create a connection to read text data directly into R functions. You need to convert FASTA sequences to DNAStringSet. You can use getSequence function

Task 3.3: Sequence Processing (20)

1. Identify sequences with gaps or ambiguous bases:
   Hint: Use the alphabetFrequency function. Try to understand what is the format of the returned value
2. Remove gaps and ambiguous bases from sequences, and determine the length of the sequence before and after the removal
3. Run Pairwise local alignment, report the score and width of each sequence before and after the alignment.
4. Create a function that returns all the positions of mismatching pairs

Reporting (5 points)

- You will be marked on reporting and code readability and cleanliness
- Write an introduction to the complete report and an introduction for each section of your code
- Write clear comments

References

- BrainCancer Dataset
- About the BrainCancer Dataset
- Principal Component Analysis in R
- Diabetes Prediction Dataset
- ggplot2 Documentation
- Fisher Test in R
- T-Test in R
- Fasta Format
- Pairwise Alignment
- S4 Object
- Markdown Cheatsheet