

TRGN 527: Applied Data Science and Bioinformatics

UNIT I. Introduction and Basic Data Science

Week 4 – Assignment 4

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | Assistant Professor

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

David W. Craig, Ph.D. | Professor and Vice Chair

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

Assignment 4

- Instructions: Please answer the following questions, first 14 questions worth 7 points and last 2 questions worth 1 point (Total: 100 points).
- Load the file “TRGN599_A4.mzXML” in your R-Studio.
 - For example, using the following code:

```
45 library(mzR)
46 library(msdata)
47 library(magrittr)
48
49 TRGN599_MS_file <- 'TRGN599_A4.mzXML'
50
51 my.data <- openMSfile(TRGN599_MS_file)
52
```

- In order to answer the following questions you should
- 1) create an R markdown file as explained during the last class,
- 2) copy/paste & run with the new data file (TRGN599_A4.mzXML) the first 129 lines of code of the case study 1 Rmarkdown file (proteomics analysis) that you can find in Blackboard/Lecture 4:
 - File name: Rmarkdown_TRGN_599_Week_4_Proteomics_Analysis.Rmd.

Assignment 4

- Question 1

- Which of the following R commands gives you access to the measurement-metadata?
 - A) `runInfo(my.data)`
 - B) `openMSfile(my.data)`
 - C) `boxplot(my.data)`
 - D) `plot(my.data)`

- Question 2

- Which is the number of actual spectra in the set by using the `runInfo(my.data)$scanCount` function?
 - A) 10
 - B) 5
 - C) 2
 - D) 7

Assignment 4

- Question 3

- Which is the ion source for the experiment by using the *“instrumentInfo(my.data)\$ionisation”* function?
 - A) hybridization ionization
 - B) fluorescence ionization
 - C) electrophoresis ionization
 - D) electrospray ionization

- Question 4

- How many columns you have in your data when use the *header(my.data)* function?
 - A) 50
 - B) 26
 - C) 31
 - D) 33

Assignment 4

- Question 5

- When you plot the spectrum using the `plot(pl, type="h")` function how many peaks you observe greater than $4e+07$?
 - A) 20
 - B) 10
 - C) 3
 - D) 9

- Question 6

- When you pick up the 15 strongest peaks using the function
 - `topnum <- 15`
 - `pl.top <- pl[pl[,2] %in% head(sort(pl[,2],decreasing=T),topnum),1]`
- Which of the following is the highest peak?
 - A) 464.8184
 - B) 924.7365
 - C) 395.8215
 - D) 789.7467

Assignment 4

- Question 7

- When you pick up the 15 strongest peaks using the function
 - `"topnum <- 15"`
 - `pl.top <- pl[pl[,2] %in% head(sort(pl[,2],decreasing=T),topnum),1]"`
- Which of the following is the lowest peak?
 - A) 464.8184
 - B) 924.7365
 - C) 395.8215
 - D) 789.7467

- Question 8

- When you pick up the 20 strongest peaks using the function
 - `"topnum <- 15"`
 - `pl.top <- pl[pl[,2] %in% head(sort(pl[,2],decreasing=T),topnum),1]"`
- Which of the following is the highest peak?
 - A) 333.6765
 - B) 924.7365
 - C) 395.8215
 - D) 789.7467

Assignment 4

- Question 9

- When you pick up the 20 strongest peaks using the function:
 - `“topnum <- 15`
 - `pl.top <- pl[pl[,2] %in% head(sort(pl[,2],decreasing=T),topnum),1]”`
- Which of the following is the lowest peak?
 - A) 333.6765
 - B) 924.7365
 - C) 395.8215
 - D) 789.7467

- Question 10

- When you plot the peak differences using the function:
 - `peakdiff <- outer(pl.top, pl.top, '-')`
 - `plot(density(abs(peakdiff)))`
- Which of the following is the probability function of the graph:
 - A) Probability density function (PDF)
 - B) Probability mass function (PMF)
 - C) Poisson distribution
 - D) Random distribution

Assignment 4

- Question 11

- When you plot the peak differences using the function:
 - `peakdiff <- outer(pl.top, pl.top, '-')`
 - `plot(density(abs(peakdiff)))`
- Which of the following would be the most suitable description of the observed curve distributions:
 - A) One can observe that the curve is a composition of two wide distributions.
 - B) One can observe that the curve is a composition of three wide distributions.
 - C) One can observe that the curve is a composition of one wide distributions.
 - D) One can observe that the curve is a composition of five wide distributions.

- Question 12

- When you plot the peak differences using the function:
 - `peakdiff <- outer(pl.top, pl.top, '-')`
 - `plot(density(abs(peakdiff)))`
- And previously used the 20 strongest peaks using the functions:
 - `topnum <- 20`
 - `pl.top <- pl[pl[,2] %in% head(sort(pl[,2],decreasing=T),topnum),1]`
 - `pl.top`
- Which of the following would be the most suitable description of the peakdiff matrix:
 - A) It is by itself a 20x20 matrix containing the differences among the elements of the pl.top vector
 - B) It is by itself a 60x60 matrix containing the differences among the elements of the pl.top vector
 - C) It is by itself a 10x10 matrix containing the differences among the elements of the pl.top vector
 - D) It is by itself a 15x15 matrix containing the differences among the elements of the pl.top vector

Assignment 4

- Question 13

- When you plot the peak differences using the function:
 - `peakdiff <- outer(pl.top, pl.top, '-')`
 - `plot(density(abs(peakdiff)))`
- And previously used the 20 strongest peaks using the functions:
 - `topnum <- 20`
 - `pl.top <- pl[pl[,2] %in% head(sort(pl[,2],decreasing=T),topnum),1]`
 - `pl.top`
- Which of the following numbers is the sample size of the plot?
 - A) N = 200
 - B) N = 300
 - C) N = 500
 - D) N = 400

Assignment 4

- Question 14

- When you plot the peak differences using the function:
 - `peakdiff <- outer(pl.top, pl.top, '-')`
 - `plot(density(abs(peakdiff)))`
- And previously used the 20 strongest peaks using the functions:
 - `topnum <- 20`
 - `pl.top <- pl[pl[,2] %in% head(sort(pl[,2],decreasing=T),topnum),1]`
 - `pl.top`
- Which of the following numbers is the bandwidth of your plot?
 - A) Bandwidth = 32.53
 - B) Bandwidth = 53.52
 - C) Bandwidth = 22.53
 - D) Bandwidth = 52.53

Assignment 4

- Question 15

- Which of the following is an analytic technique that produces mass-to-charge ratio spectra of peptide fragment in samples:
 - A) Microarrays
 - B) Mass spectrometry (MS)
 - C) Messenger RNA (mRNA)
 - D) Proteomics

- Question 16

- Regarding the different file formats for MS data, which file format you are using in your proteomic analysis when uploading data through the mzR package from Bioconductor?
 - A) mzML
 - B) netCDF
 - C) mzXML
 - D) mzCDF