Problemset 2

Malay Basu (malay@uab.edu)

Problem 1

You can call `data(airquality)` in R. It will generate the following data.

```
data("airquality")
knitr::kable(head(airquality))
```

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6

(a) Calculate the pairwise Pearson correlation of all the variables of this dataset and create a dataframe that has columns like this.

var1	var2	corr
Ozone	Solar	0.4

(b) Using `ggplot` draw the scatterplot of the variables that show the highest correlation. You can arbitrarily choose one of the two variables as independent. Make the scatterplot publication quality. Also calculate the r.sq of the plot and put it on the top of the plot as subtitle of the plot.

Problem 2

Write an R script that takes two arguments: (1) a fasta file name, (2) a sequence ID. The script should print out the sequence matching the id in FASTA format to the terminal.

Problem 3

Using `wget` download BLOSUM62 matrix from NCBI FTP server (ftp://ftp.ncbi.nih.gov/blast/matrices/BLOSUM62). Process it in anyway you can and read it in R as a matrix, a dataframe, or a list. You should store the data such a way that you can call the score given two amino acids as key as a fast lookup table. Read the accompanied `ex_align.fas` file and calculate the score of the given alignment. Consider each indel has score 0. The alignment file is in aligned fasta format.

Tips: You need to use either `seqnir` or `Biostrings` package and loop through each position in the alignment.