A3,712,2022

Chen, Liyang

2022-05-27

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

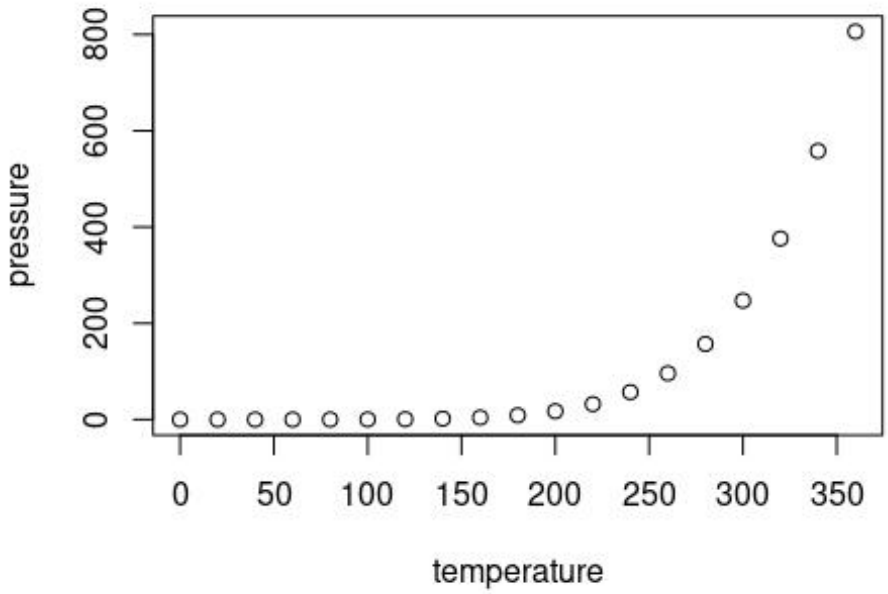
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Assessment 3"

Part 1-1 Description of the file gene_expression.tsv"

This report was completed by a single person, thus this report does not include Team Coding Evidence that describes the team members.

Question 1: Read in the file, making the gene identifiers the row names. Show a table of values for the first six genes.

Representing the first six values of the raw data in a table "Q1_table". The data is added into "Origin"

url <-
"https://raw.githubusercontent.com/markziemann/SLE712_files/master/assessment_task3/bioinfo_asst3_part1_files/gene_expression.tsv"
download.file(url,destfile="gene_expression") Origin<-read.table(url) View(Origin)
Origin<-read.table(url,stringsAsFactors=FALSE, header=TRUE, row.names=1)
head(Origin) rownames(Origin) grep("ENG",rownames(Origin)) Origin[1:6,]
Q1_table <- Origin[1:6,] head(Q1_table, n=6) View(Q1_table)

Question 2: Make a new column which is the mean of the other columns. Show a table of values for the first six genes.

A new column is created which is the average of the other columns, and it is add in Table"Origin". A new table"Q2_table" includes the first six data with added column.

ncol(Origin) rowMeans(Origin) Mean<-rowMeans(Origin) Origin$Mean<-Mean head(Origin) Origin[1:6,] Q2_table <- Origin[1:6,] head(Q2_table, n=6) View(Q2_table)

Question 3: List the 10 genes with the highest mean expression

The Order command can be used to filter the desired value based on the results of an Internet search. Q3_table is used to accommodate the filtered values of the 10 largest average values

T1 <- Origin[order(-Origin$Mean),] Q3_table <- head(T1, n=10) head(Q3_table, n=10) View(Q3_table)

Question 4: Determine the number of genes with a mean <10

Number_10 is used to record values with an average value less than 10.

Number_10 <- subset(Origin, Origin$new_column < 10) nrow(Number_10)

The final result is 0.

The subset command can filter the data that meets the requirements in the collection or table, so in this question, the values that meet the "average less than 10" condition can be filtered

Question 5: Make a histogram plot of the mean values and include it into your report

hist(Origin$Mean, xlab = "Mean values",main = "Mean")

Making diagrams refered to the practical 1 on April 10, 2022. The course provided methods for drawing diagrams in Rstudio.

Part 1-2 Description of the file growth_data.csv

Question 6: Import this csv file into an R object. What are the column names?

Origin2 is the code name for the total data in this section url_part is the code of growth_data.csv in this question.

After the data is downloaded from the web page, useing str command and head command compress and filter the redundant information that is not needed for this question.

url_part <- "https://raw.githubusercontent.com/markziemann/SLE712_files/master/assessment_task3/bioinfo_asst3_part1_files/growth_data.csv" download.file(url_part, destfile = "growth_data.csv") Origin2 <-read.csv("growth_data.csv", header= TRUE, stringsAsFactors =FALSE) str(Origin2) head(Origin2) View(Origin2)

Question 7: Calculate the mean and standard deviation of tree circumference at the start and end of the study at both sites.

In this question, the mean and standard deviation of the two periods need to be calculated separately.

The mean and standard deviation in Northeast

Data_NE is data of trees in Northeast M_NE is Mean of Northeast SD_NE is Standard Deviation of Northeast

Data_NE <- subset(Origin2,Site=="northeast") head(Data_NE) tail(Data_NE) str(Data_NE) M_NE_2005 <- mean(Data_NE$Circumf_{2005_c}m$)M_NE_{2020} <− $mean(Data_NE$Circumf_2020_cm) SD_NE_2005 <- sd(Data_NE$Circumf_{2005_c}m$)SD_NE_{2020} <− $sd(Data_NE$Circumf_2020_cm)

M_NE_2005 is 5.292, M_NE_2020 is 54.228 SD_NE_2005 is 0.9140, SD_NE_2020 is 25.2279

The mean and standard deviation in Southwest

Data_SW is data of trees in Southwest M_SW is Mean of Southwest SD_SW is Standard Deviation of Southwest

Data_SW <- subset(Origin2,Site=="southwest") head(Data_SW) tail(Data_SW) str(Data_SW) M_SW_2005 <- mean(Data_SW$Circumf_{2005_c}m$)M_SW_{2020} <− $mean(Data_SW$Circumf_2020_cm) SD_SW_2005 <- sd(Data_SW$Circumf_{2005_c}m$)SD_SW_{2020} <− $sd(Data_SW$Circumf_2020_cm)

M_SW_2005 is 4.862, M_SW_2020 is 45.596 SD_SW_2005 is 1.1475, SD_SW_2020 is 17.8735

Question 8: Make a box plot of tree circumference at the start and end of the study at both sites.

According to the Practical 1 (Bioinformatics Prac 1 - Hello World!), this problem is completed by a simple programming.

1. The box polt of the Northeast

Data_NE is tree data of northeast from Question 7

x_vals <- Data_NE$Circumf_{2005_c} my_v als <- Data_N E$Circumf_2020_cm

boxplot(x_vals, y_vals, main = "The box plot of tree circumference (Northeast)", xlab = "Year (2005-2020)", ylab = "Tree Circumference (cm)", names = c("2005", "2020"), col = "green")

2. The box polt of the Southwest

Data_SW is tree data of southwest from Question 7

x_vals <- Data_SW$Circumf_{2005_c} my_v als <- Data_S W$Circumf_2020_cm

boxplot(x_vals, y_vals, main = "The box plot of tree circumference (Southwest)", xlab = "Year (2005-2020)", ylab = "Tree Circumference (cm)", names = c("2005", "2020"), col = "yellow")

Question 9: Calculate the mean growth over the last 10 years at each site.

Northeast Mean_10_NE is the mean growth in northeast

Data_NE$growth$ <- $Data_N E$Circumf_2020_cm-Data_NE$Circumf_{2010_c} mhead(Data_N E) Mean_{10_N} E$ <- $mean(Data_N E$growth)
View(Mean_10_NE)

Mean_10_NE is 42.94

Southwest Mean_10_SW is the mean growth in northeast

Data_SW$growth$ <- $Data_S W$Circumf_2020_cm-Data_SW$Circumf_{2010_c} mhead(Data_S W) Mean_{10_S} W$ <- $mean(Data_S W$growth)
View(Mean_10_SW)

Mean_10_SW is 35.49

Question 10: Use the t.test and wilcox.test functions to estimate the p-value that the 10 year growth is different at the two sites.

Northeast t.test(Data_NE$growth$)$wilcox.test(Data_NE$growth$)

p-value of northeast is 7.79e-10

Southwest t.test(Data_SW$growth$)$wilcox.test(Data_SW$growth$)

p-value of southwest is 7.782e-10

Part 2: Examining biological sequence diversity

Question 2-1: Download the whole set of coding DNA sequences for E. coli and your organism of interest. How many coding sequences are present in these organisms? How much coding DNA is there in total for these two organisms? Describe any differences between the two organisms.

Download E. coli data.

Ecoli is a dataset of Escherichia coli

library("R.utils") url_Ecoli <- "http://ftp.ensemblgenomes.org/pub/bacteria/release-53/fasta/bacteria_0_collection/escherichia_coli_str_k_12_substr_mg1655_gca_000005845/cds/Escherichia_coli_str_k_12_substr_mg1655_gca_000005845.ASM584v2.cds.all.fa.gz" download.file(url_Ecoli,destfile="ecoli_cds.fa.gz")

library("seqinr") Ecoli <- seqinr::read.fasta("ecoli_cds.fa") str(head(Ecoli)) length(Ecoli) head(summary(Ecoli)) count(cds[[1]],3) sum(sapply(cds[1:3], length))

Coding Sequences of E.coli is 3462

Download Required Organism The required organism is Salmonella enterica subsp. enterica serovar Weltevreden (GCA_005518735).

W is a dataset of enterica serovar Weltevreden

library("R.utils") url_W <- "http://ftp.ensemblgenomes.org/pub/bacteria/release-53/fasta/bacteria_50_collection/salmonella_enterica_subsp_enterica_serovar_weltevreden_gca_005518735/cds/Salmonella_enterica_subsp_enterica_serovar_weltevreden_gca_005518735.ASM551873v1.cds.all.fa.gz" download.file(url_W,destfile="weltevreden_cds.fa.gz") gunzip("weltevreden_cds.fa.gz")

library("seqinr") W <- seqinr::read.fasta("weltevreden_cds.fa") str(head(W)) length(W) head(summary(W)) count(cds[[1]],3) sum(sapply(cds[1:3], length))

Coding Sequences of Weltevreden is 3462

Question 2-2: Calculate the length of all coding sequences in these two organisms. Make a boxplot of coding sequence length in these organisms. What is the mean and median coding sequence length of these two organisms? Describe any differences between the two organisms.

1. Calculation

E.coli Ecoli is the data of E.coli from Question 2 Lenth_Ecoli is used for the lenth of E.coli coding

head(summary(Ecoli)[,1]) Lenth_Ecoli <- as.numeric(summary(Ecoli)[,1]) sum(Lenth_Ecoli)

Lenth_Ecoli is 3978528

Weltevreden W is the data of Weltevreden from Question 2 Lenth_W is used for the lenth of Weltevreden coding

head(summary(W)[,1]) Lenth_W <- as.numeric(summary(W)[,1]) sum(Lenth_W)

Lenth_W is 4278831

2. Boxpolt, Mean and Median

In this question, the boxplot, mean and median need to be represented separately

E.coli M_E is the mean of E.coli, and M_ME is the median

boxplot(Lenth_Ecoli,ylab="sequence length (bp)") M_E <- mean(Lenth_Ecoli) M_ME <- median(Lenth_Ecoli)

M_E is 938.5534, and M_ME is 831

Weltevreden W_E is the mean of Weltevreden, and W_ME is the median

boxplot(Lenth_W,ylab="sequence length (bp)") W_E <- mean(Lenth_W) W_ME <- median(Lenth_W)

W_E is 915.06223, W_ME is 786

Question 2-3: Calculate the frequency of DNA bases in the total coding sequences for both organisms. Perform the same calculation for the total protein sequence. Create bar plots for nucleotide and amino acid frequency. Describe any differences between the two organisms.