Solutions to Module 2 Homework

Problem 1 (30 points)

Complete the following computations on gene means of the Golub data set.

a) Compute the mean expression values for every gene among “ALL” patients.
b) Compute the mean expression values for every gene among “AML” patients.
c) Give the biological names of the three genes with the largest mean expression value among “ALL” patients.
d) Give the biological names of the three genes with the largest mean expression value among “AML” patients.

Answer:

```
# Solutions to Module 2 Homework

#Complete the following computations on gene means of the Golub data set.
# a) Compute the mean expression values for every gene among “ALL” patients.
rm(list = ls())

data(golub)
library(multtest)
gol.fac<-factor(golub.cl, levels=0:1, labels=c("ALL","AML"))
meanALL<-apply(golub[,gol.fac=="ALL"], 1, mean)
meanALL

# b) Compute the mean expression values for every gene among “AML” patients

data(golub)
library(multtest)
gol.fac<-factor(golub.cl, levels=0:1, labels=c("ALL","AML"))
meanAML<-apply(golub[,gol.fac=="AML"], 1, mean)
meanAML
```

Solutions to Module 2 Homework

c) Give the biological names of the three genes with the largest mean expression value among “ALL” patients.

```
data(golub)
library(multtest)
gol.fac<-factor(golub.cl, levels=0:1, labels=c("ALL","AML"))
meanALL<-apply(golub[,gol.fac=="ALL"], 1, mean)
orderALL<-order(meanALL, decreasing=TRUE)
golub.gnames[orderALL[1:3],2]

[1] "GB DEF = Chromosome 1q subtelomeric sequence D1S553"
[2] "37 kD laminin receptor precursor/p40 ribosome associated protein gene"
[3] "RPS14 gene (ribosomal protein S14) extracted from Human ribosomal protein S14 gene"
>
```

d) Give the biological names of the three genes with the largest mean expression value among “AML” patients.

```
data(golub)
library(multtest)
gol.fac<-factor(golub.cl, levels=0:1, labels=c("ALL","AML"))
meanAML<-apply(golub[,gol.fac=="AML"], 1, mean)
orderAML<-order(meanAML, decreasing=TRUE)
golub.gnames[orderAML[1:3],2]

[1] "GB DEF = mRNA fragment for elongation factor TU (N-terminus)"
[2] "GB DEF = HLA-B null allele mRNA"
[3] "Globin, Beta"
```

Solutions to Module 2 Homework

Problem 2 (30 points)

a) Save the expression values of the first five genes (in the first five rows) for the AML patients in a csv file “AML5.csv.”
b) Save the expression values of the first five genes for the ALL patients in a
plain text file “ALL5.txt.”
c) Compute the standard deviation of the expression values on the first patient, Of the 100th to 200th genes (total 101 genes).
d) Compute the standard deviation of the expression values of every gene, across all patients. Find the number of genes with standard deviations greater than 1.
e) Do a scatter plot of the 101st gene expressions against the 102nd gene expressions, labeling the x-axis and the y-axis with the genes’ biological names. Do this using xlab= and ylab= control options.

Answer:

a) Save the expression values of the first five genes (in the first five rows) for the AML patients in a csv file “AML5.csv.”

```
data(golub)
library(multtest)
gol.fac<-factor(golub.cl, levels=0:1, labels=c("ALL","AML"))
golub.AML<-golub[,gol.fac=="AML"]
AML5<-golub.AML[1:5,]
getwd()
write.csv(AML5,file="AML5.cv")
```

Solutions to Module 2 Homework

b) Save the expression values of the first five genes for the ALL patients in a plain text file “ALL5.txt.”

```
data(golub)
library(multtest)
gol.fac<-factor(golub.cl, levels=0:1, labels=c("ALL","AML"))
golub.ALL<-golub[,gol.fac=="ALL"]
ALL5<-golub.ALL[1:5,]
getwd()
write.table(ALL5,file="ALL5.txt")
```

c) Compute the standard deviation of the expression values on the first patient, of the 100th to 200th genes (total 101 genes).

```
> data(golub)
> library(multtest)
> sd<-sd(golub[100:200,1])
> sd
[1] 0.9174976
```

d) Compute the standard deviation of the expression values of every gene, across all patients. Find the number of genes with standard deviations greater than 1.

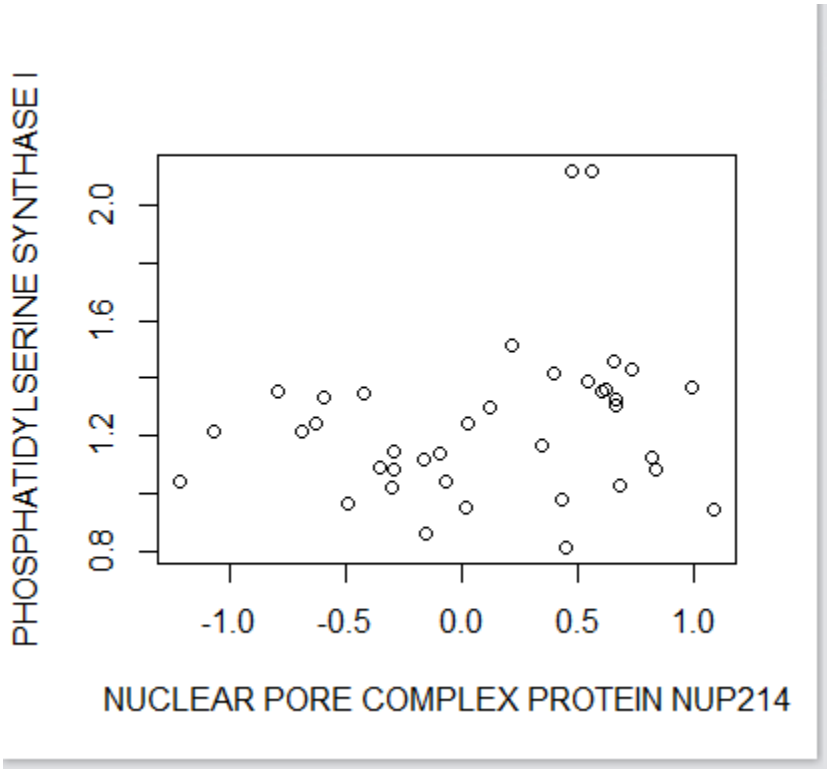
```
> data(golub)
> library(multtest)
> sd.all<-apply(golub, 1,sd)
```

Solutions to Module 2 Homework

```
> print(sum(sd.all>1))
[1] 123
```

e) Do a scatter plot of the 101st gene expressions against the 102nd gene expressions, labeling the x-axis and the y-axis with the genes' biological names. Do this using xlab= and ylab= control options.

```
data(golub)
library(multtest)
plot(golub[101,],golub[102,],xlab=golub.gnames[101,2],
ylab=golub.gnames[102,2])
```



Solutions to Module 2 Homework

Problem 3 (20 points)

Complete a–c using the ALL data set.

Load the ALL data from the ALL library, and use str and openVignette() for further orientation.

a) Use exprs(ALL[,ALL$BT=="B1"] to extract the gene expressions from the patients in disease stage B1. Produce a histogram of these gene expressions.

b) Compute the mean gene expressions over these B1 patients.

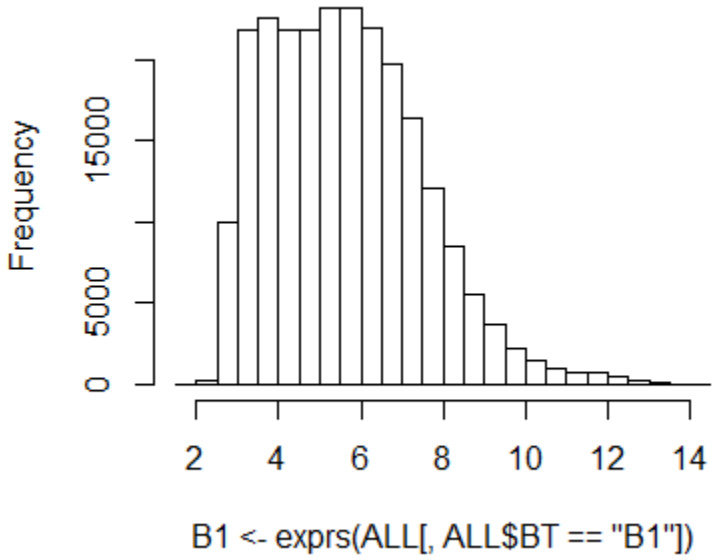
c) Give the gene identifiers of the three genes with the largest mean. Submit R commands that fulfill the tasks in a-c, and answer part c directly.

Solutions to Module 2 Homework

Answer:

a) Use exprs(ALL[,ALL$BT=="B1"] to extract the gene expressions from the patients in disease stage B1. Produce a histogram of these gene expressions.

```
data(ALL)
library(ALL)
hist(B1<-exprs(ALL[,ALL$BT=="B1"]))
```



b) Compute the mean gene expressions over these B1 patients

```
data(ALL)
library(ALL)
B1<-exprs(ALL[,ALL$BT=="B1"])
mean.B1<-apply(B1,1,mean)
```

```
mean.B1
```

c) Give the gene identifiers of the three genes with the largest mean.

```
>
> data(ALL)
> library(ALL)
> B1<-exprs(ALL[,ALL$BT=="BT"])
> order.B1<-order(mean.B1,decreasing=TRUE)
> mean.B1[order.B1[1:3]]
AFFX-hum_alu_at       31962_at      31957_r_at
      13.41648       13.16671       13.15995
>
```

These are the three genes with the largest mean

- AFFX-hum_alu_at
- 31962_at
- 31957_r_at

Problem 4 (20 points)

To complete a and b, work with the “trees” data set that comes with R.

a) Find the type of the trees data object.

b) Produce a figure with two overlaid scatterplots: girth versus height and girth versus volume. Do the height plot with blue “+” symbols, and do the volume plot with red “o” symbols. You need to set the ylim= control option so that all points from the two plots can show up on the merged figure.

Solutions to Module 2 Homework

Answer:

a) Find the type of the trees data object.

```
> class(trees)
[1] "data.frame"
>
```

b) Produce a figure with two overlaid scatterplots: girth versus height and girth versus volume. Do the height plot with blue "+" symbols, and do the volume plot with red "o" symbols. You need to set the ylim= control option so that all points from the two plots can show up on the merged figure.

```
rm(list=ls())

str(trees)

plot(trees$Height~trees$Girth,col="blue",pch="+",xlim=c(10,18),ylim=c(0,100),xlab="Girth",ylab="Height and Volume")

points(trees$Girth,trees$Volume,col="red",pch="O")

legend("bottomright", c("Height","volume"), col=c("blue","red"), pch=c("+","O"))
```

Solutions to Module 2 Homework