Bioinformatics Lab5

Saman Zahid, Milda Poceviciute, Fanny Karelius, Rab Nawaz Jan Sher

December 11, 2018

Question 1

```
#install.packages("devtools")
library(devtools)
#install_github("mategarb/R.ROSETTA")
library(R.ROSETTA)
```

Question 2

Total number of features = 36 - 1 = 35, since last one is class.

Number of objects in each class.

```
##
##  autism control
##      82      64
```

Is distribution balanced?

```
control_ratio <- as.numeric(decision_sum[1])/146
autism_ratio <- as.numeric(decision_sum[2])/146
```

Since controls and autism are nearly around 50% distributed thus the distribution is balanced

```
autconDefault = rosetta(autcon)

rules <- autconDefault$main
quality <- autconDefault$quality
quality
```

```
##  Accuracy.Mean Accuracy.Median Accuracy.Std Accuracy.Min Accuracy.Max
##       0.821818             0.8     0.083158     0.733333            1
```

Part a - Cross validation: Partition the data into equally sized subsets and perform model assesment.

Rosetta performs 10 fold cross validation by default

Part b - "Johnson" is the default reduction method. Reducer is used to find the dependencies in data.

Part c - The default method of discretization is **EqualFrequency**. It partitions the range of features a(x) into (n+1) intervals by placing cuts between features, so that each interval has the same number of objects.

The default number of bins is 3.

Part d

```
paste("Accuracy of model : " , quality$Accuracy.Mean)
```

```
## [1] "Accuracy of model :  0.821818"
```

Part e

```r
paste("Total number of rules : " ,nrow(rules))
```

```
## [1] "Total number of rules :  191"
```

```r
paste("3 Most significant rules are:")
```

```
## [1] "3 Most significant rules are:"
```

```r
rules[(1:3),]
```

```
##            FEATURES DECISION              CUTS_COND DISC_CLASSES SUPP_LHS
## 1 NCKAP5L,234817_at  control      value<cut,value<cut          1,1       18
## 2     MAP7,ATXN8OS  control      value>cut,value<cut          3,1       18
## 3     ZSCAN18,NPR2  control value<cut,cut<value<cut          1,2       19
##   SUPP_RHS ACC_RHS COV_LHS COV_RHS STAB_LHS STAB_RHS   CUT_1   CUT_2
## 1       18 0.97368 0.13740 0.30196        1        1 1.90584 1.64213
## 2       18 1.00000 0.13308 0.29932        1        1 2.51985 2.22742
## 3       19 0.98521 0.14616 0.32895        1        1 2.35647 2.54040
##     CUT_3 CUT_4         PVAL   RISK_PVAL REL_RISK    CONF_INT
## 1     NaN   NaN 4.818175e-06 0.005285147  2.28125 1.273:4.089
## 2     NaN   NaN 4.818175e-06 0.005285147  2.28125 1.273:4.089
## 3 2.59265   NaN 4.818175e-06 0.003949585  2.28125 1.298:4.009
```

```r
ind <- rules$PVAL < 0.05
sig_rules <- rules$DECISION[ind]
class_control <- length(which(sig_rules == "control"))
class_autism <- length(which(sig_rules == "autism"))

paste("Number of significant rules for class control : " ,class_control)
```

```
## [1] "Number of significant rules for class control :  77"
```

```r
paste("Number of significant rules for class autism : " ,class_autism)
```
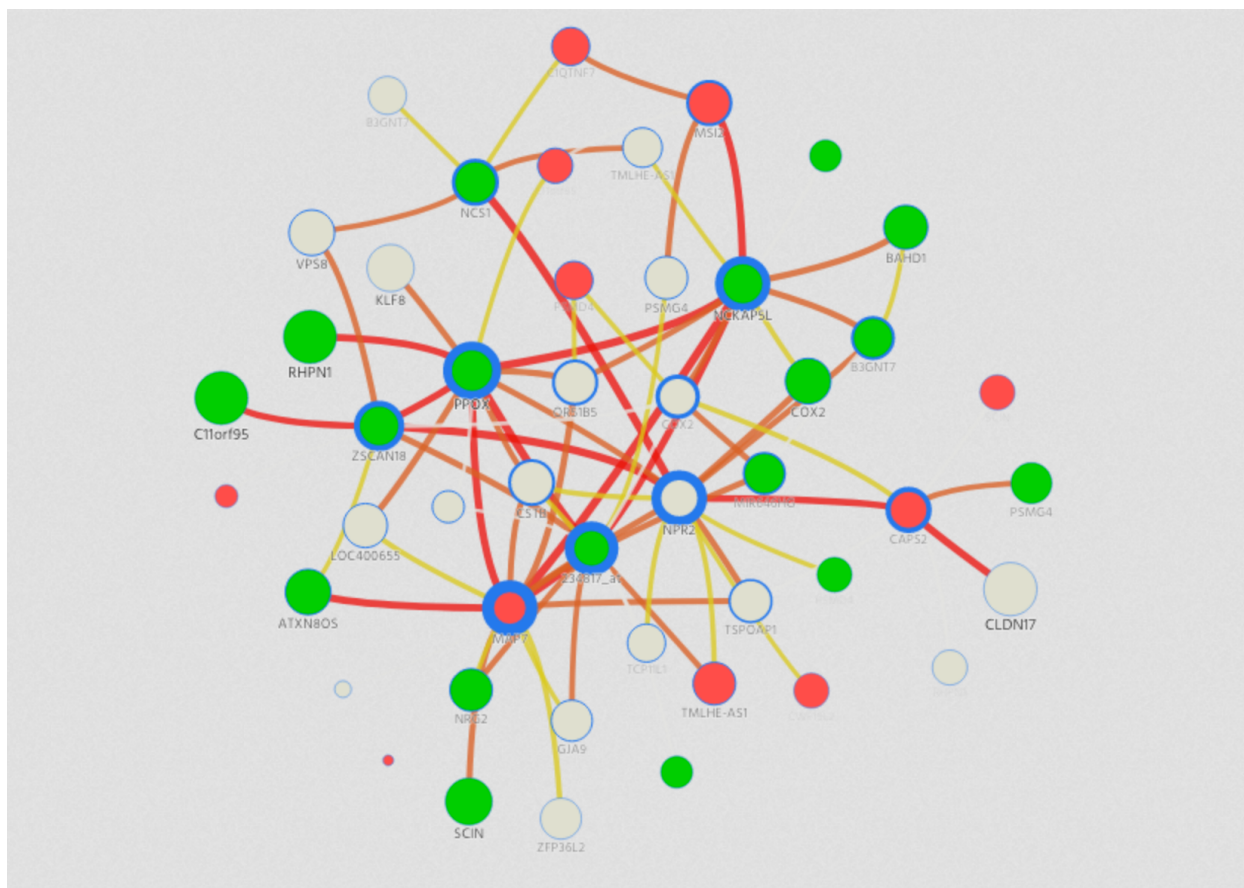
```
## [1] "Number of significant rules for class autism :  108"
```

Thus autism gets more significant rules.
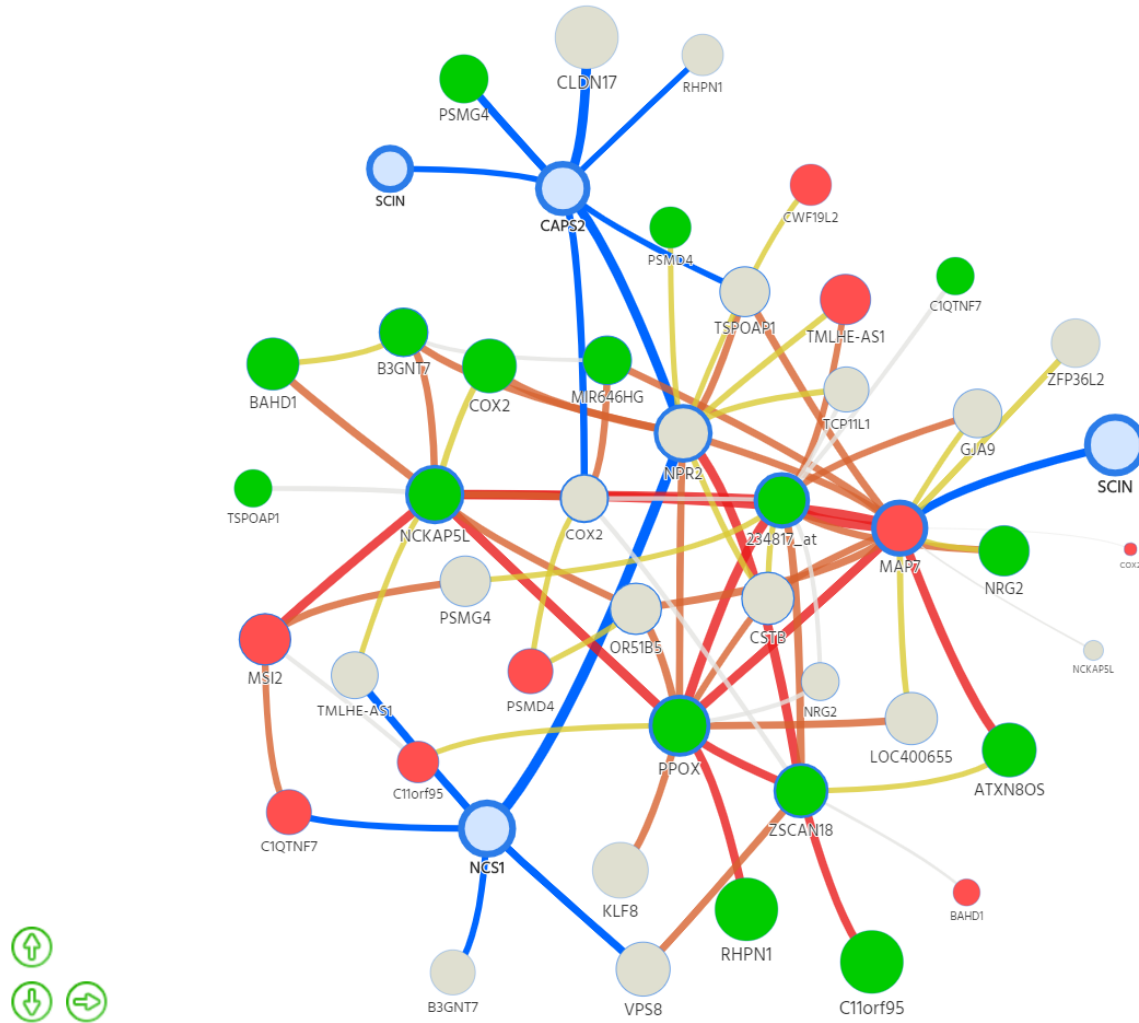
Question 5 and 6

Control Network

```r
knitr::include_graphics('images/control.png')
```

Autism Network

```
knitr::include_graphics('images/autism.png')
```

Controls: There are 80 rules in decision "control". The most significant nodes with highest number of connections are "MAP7", "PPOX" , "NPR2","234817_at" and "NCKAP5L". Out of these significant nodes, "MAP7" is overexpressed and "PPOX", "234817_at", "NCKAP5L" are underexpressed.

Autism: There are 111 rules in deciion "autism". The most significant nodes with highest number of connections are "COX2","RHPN1","PSMG4","SCIN","MAP7". It should be noticed that "MAP7" is either underexpressed or have no change in autism while in control it was overexpressed.

```
knitr::include_graphics('images/autism_caliaum_ion.PNG')
```

```
knitr::include_graphics('images/control_calcium_ion.PNG')
```

Calcium dependent gene expression regulates brain development and if it is compromised or develops abnormally, it might leads to autism. It can be observed that the calcium ion binding genes "SCIN","NCS1" and "CAPS2" have more densed network of connection in autism as compared to control.

SAFARI database contains information about autism related genes. It can be seen that we have "COX2" and "TMLHE-AS1" appears in our autism network which are also identified in SAFARI database symbolized as "CUX2" and "TMLHE".

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
#install.packages("devtools")
library(devtools)
#install_github("mategarb/R.ROSETTA")
library(R.ROSETTA)


#print(autcon)
```

```r
decision_sum <- table(autcon$decision)
decision_sum


control_ratio <- as.numeric(decision_sum[1])/146
autism_ratio <- as.numeric(decision_sum[2])/146

autconDefault = rosetta(autcon)

rules <- autconDefault$main
quality <- autconDefault$quality
quality

paste("Accuracy of model : " , quality$Accuracy.Mean)

paste("Total number of rules : " ,nrow(rules))
paste("3 Most significant rules are:")
rules[(1:3),]



ind <- rules$PVAL < 0.05
sig_rules <- rules$DECISION[ind]
class_control <- length(which(sig_rules == "control"))
class_autism <- length(which(sig_rules == "autism"))

paste("Number of significant rules for class control : " ,class_control)

paste("Number of significant rules for class autism : " ,class_autism)
knitr::include_graphics('images/control.png')
knitr::include_graphics('images/autism.png')
knitr::include_graphics('images/autism_caliaum_ion.PNG')
knitr::include_graphics('images/control_calcium_ion.PNG')
```