Bioinformatics - Computer Lab 5

Group 7: Lennart Schilling (lensc874), Thijs Quast (thiqu264), Mariano Maquieira Mariani (marma330)

13-12-2018

Contents

Assignment 1

```
library(devtools)
library(R.ROSETTA)
```

Assignment 2

```
data(autcon)
```

```
dim(autcon)
```

```
## [1] 146  36
```

The "autcon" dataset has 36 - 1 = 35 features for 146 observations.

```
table(autcon$decision)
```

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

```
percentage_autism <- 82/146
percentage_control <- 64/146
percentage_autism
```

```
## [1] 0.5616438
```

```
percentage_control
```

```
## [1] 0.4383562
```

In total 82 decisions result in autism, whereas 64 decisions result in control. Therefore one can say the dataset is approximately in balance.

Assignment 3

```
autconDefault = rosetta(autcon)
table <- autconDefault$main
#save(list = ls(all = TRUE), file = "outcome.Rdata")
```

```
load("outcome.Rdata")
autconDefault$quality
```

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

3a

Cross validation means that the dataset is divided into k different folds. Then one fold is extracted from the dataset and functions as test dataset. The model is than run k different times with k different folds as test datasets. Model parameters are then estimated by taking an average, or confidence interval from the output of these k number of models. By default, the rosetta function uses 10 folds.

3b

The default reduction method is “Johnson”. The reduction method creates a minimal subset of attributes so that it preserves indiscernability between objects. This is useful so that an optimal subset of features can be extracted.

3c

The default method for discretization is EqualFrequency. Discretization subdivides e.g. continuos data into different classes. E.g. temperatues from different ranges are named as “Hypothermia”, “Fever”, “Normal”, “Hyperthermia”.Equal frequency

3d

The accuracy of the model approximately 82.18% the mean, and approximately 76.36% accuracy on the median.

3e

```
table <-table[order(table$PVAL),]
top_three <- table[1:3, 1]
top_three
```

```
## [1] "NCKAP5L,234817_at" "MAP7,ATXN8OS"      "ZSCAN18,NPR2"
```

```
subset_significant <- subset(table, table$PVAL < 0.05)
subset_autism <- subset(subset_significant, subset_significant$DECISION == "control")
subset_control <- subset(subset_significant, subset_significant$DECISION == "autism")

significant_autism <- nrow(subset_autism)
significant_control <- nrow(subset_control)

significant_autism
```

```
## [1] 77
```

```
significant_control
```

```
## [1] 108
```

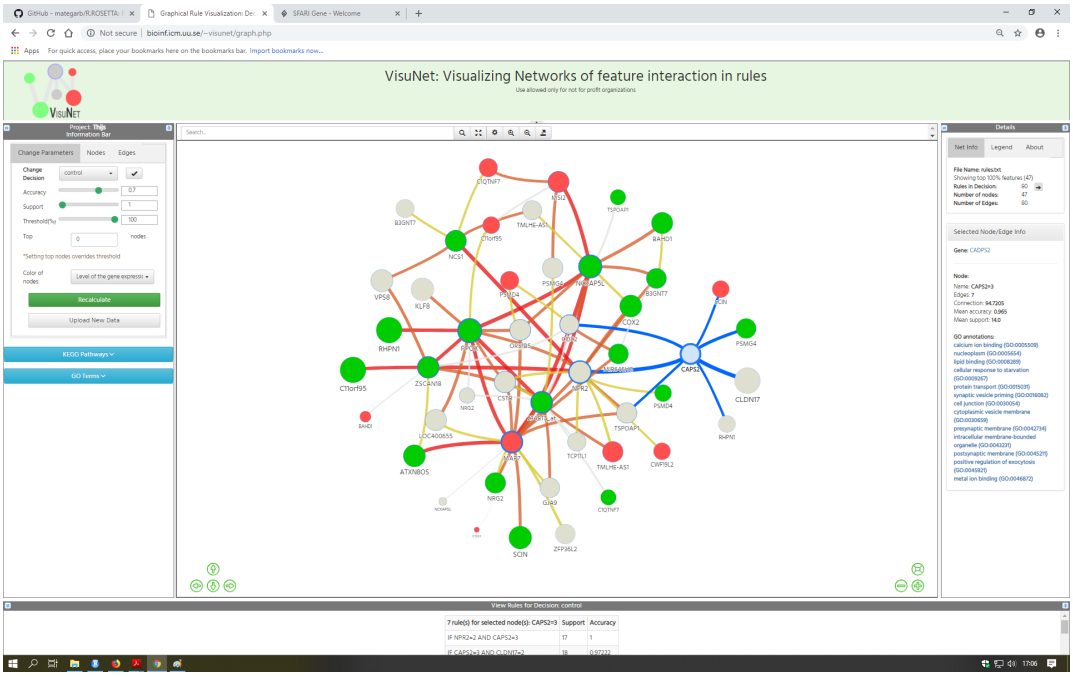
The class control gets most significant rules.

Assignment 4

```
text <- saveLineByLine(rules = table, "rules.txt")
```

Assignment 5

```
knitr::include_graphics("screenshot.png")
```



Assignment 6

The strongest connections are the red ones. The names of these are: SCIN, CAPS2, CWF19L2, TMLHE-A51, MAP7, COX2, BAHD1, PSMD4, C11orf95, C1QTNF7, MS12.

As the most significant we choose the ones with the most connections. From the left-hand side menu under "Nodes", we extract the top 5.

Node: Name: MAP7=3 Edges: 16 Connection: 224.4915 Mean accuracy: 0.931 Mean support: 14.875

Node: Name: NPR2=2 Edges: 13 Connection: 193.54575 Mean accuracy: 0.953 Mean support: 15.615

Node: Name: PPOX=1 Edges: 12 Connection: 188.37916 Mean accuracy: 0.941 Mean support: 16.667

Node: Name: 234817_at=1 Edges: 13 Connection: 180.16338 Mean accuracy: 0.919 Mean support: 15.0

Node: Name: NCKAP5L=1 Edges: 11 Connection: 171.07155 Mean accuracy: 0.973 Mean support: 16.0

For the mentioned and most significant genes, the gene ontlogies are analysed.

Most sifnificant genes

MAP7

Official name: Microtubule Associated Protein 7

This gene is responsible for the production of a microtubule-associated protein that is predominantly expressed in cells of epithelial origin.

Gene Ontology (GO) terms:

GO_ID	Qualified_GO_Term
GO:0005102	signaling receptor binding
GO:0005198	structural molecule activity
GO:0005515	protein binding

NPR2

Official name: Natriuretic Peptide Receptor 2

This gene is responsible for the production of the natriuretic peptide receptor B which is an integral membrane receptors for natriuretic peptides.

Gene Ontology (GO) terms:

GO_ID	Qualified_GO_Term
GO:0004016	adenylate cyclase activity
GO:0004383	guanylate cyclase activity
GO:0004672	protein kinase activity
GO:0005515	protein binding
GO:0005524	ATP binding

PPOX

Official name: Protoporphyrinogen Oxidase

This gene is responsible for the production of the penultimate enzyme of heme biosynthesis.

Gene Ontology (GO) terms:

GO_ID	Qualified_GO_Term
GO:0004729	oxygen-dependent protoporphyrinogen oxidase activity
GO:0016491	oxidoreductase activity
GO:0050660	flavin adenine dinucleotide binding

NCKAP5L

Official name: NCK Associated Protein 5 Like

This gene is responsible for the encoding of a protein which is regulates microtubule organization and stabilization.

Gene Ontology (GO) terms:

GO_ID	Qualified_GO_Term
GO:0005515	protein binding

Additional genes

SCIN

Official name: Scinderin

Summary: SCIN is gene, which is Ca(2+)-dependent actin-severing and is also a -capping protein.

Gene Ontology (GO) terms:

GO_ID	Qualified_GO_Term
GO:0001786	phosphatidylserine binding
GO:0003779	actin binding
GO:0005509	calcium ion binding

GO_ID	Qualified_GO_Term
GO:0005545	1-phosphatidylinositol binding
GO:0005546	phosphatidylinositol-4,5-bisphosphate binding

NCS1

Official name: Neuronal Calcium Sensor 1

Summary: The NCS1 gene is part of the neuronal calcium sensor gene family. This family encodes calcium-binding proteins which are mainly expressed in neurons. The function of the protein encoded is that it regulates G protein-coupled receptor phosphorylation in a calcium-dependent manner and can replace for calmodulin.

Gene Ontology (GO) terms:

GO_ID	Qualified_GO_Term
GO:0000287	magnesium ion binding
GO:0005245	voltage-gated calcium channel activity
GO:0005509	calcium ion binding
GO:0005515	protein binding
GO:0019901	protein kinase binding

CAPS2

Official name: Calcyphosine 2

Summary: The function of the Calcyphosine-2 is that it is a calcium-binding protein with 2 EF-hand motifs.

Gene Ontology (GO) terms:

GO_ID	Qualified_GO_Term
GO:0005432	calcium:sodium antiporter activity
GO:0005509	calcium ion binding
GO:0046872	metal ion binding

Analysis: Within the GO analysis of all three obtained Genes, all three genes show the “GO:0005509”, which takes care of calcium ion binding.

GO:0005509

The molecular function of the GO:0005509 is to interact selectively and non-covalently with calcium ions (Ca2+).

Shown below is a Ancestor chart for the GO:0005509:

Ancestor chart

Ancestor chart for GO:0005509 Chart options