Assignment 4 – Visualization and interpretation of the rule-based classifier

VT26

Introduction

csVisuNet is an interactive tool designed for the structural analysis of complex rule-based classifiers. It can be applied to any classification problem and is particularly useful for complex health-related decision tasks. The rule networks generated by csVisuNet clearly identify the interactions among key disease-related features, such as genes, metabolites, methylation sites, and their associated biological conditions.

We recently updated our package to incorporate Cytoscape, a more widely recognized bioinformatics tool for visualizing and exploring biological networks. csVisuNet is implemented in R and utilizes Cytoscape to display co-predictive networks. This tool not only provides visualization and customization options for networks derived from rule-based models but also offers a more stable interactive session. Users can annotate genes with references to popular publications, databases, and web searches accessible through an external links menu.

For the lab, you will work with the autism-control dataset, which includes gene expression measures for 82 autistic and 64 healthy (control) young males (Alter et al., 2011).

During the lab, we will create a rule-based model and use csVisuNet to construct a rule network. Additionally, we will focus on interpreting the model using various approaches, such as database searching and Gene Ontology (GO) annotations.

Previous Version

The previous version of csVisuNet called VisuNet was implemented using R and Shiny Gadgets. It included features for constructing, filtering, visualizing, and customizing networks from rule-based models. To learn more about it, visit the tutorials section at (https://komorowskilab.github.io/VisuNet/about.html) and refer to the documentation by typing `?visunet` for guidance on running and using VisuNet, the R implementation. Remember to click `Run` when the interface appears in your browser.

VisuNet installation

```
library(devtools)
devtools::install_github("komorowskilab/csVisuNet")
```

Load the VisuNet package:

```
library(csVisuNet)
```

Please note that the R.ROSETTA package is loaded automatically with csVisuNet.

Additional softwares and packages:

```
# Cytoscape
https://cytoscape.org/download.html

# Connection to Cytoscape through R

if(!"RCy3" %in% installed.packages()){
    install.packages("BiocManager")
    BiocManager::install("RCy3")
}

# GO annotation
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("clusterProfiler")

if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("org.Hs.eg.db")

#word cloud
install.packages("tm")
install.packages("SnowballC")
install.packages("wordcloud")
install.packages("RColorBrewer")
```

Tasks

1. Read the `autism_control.RDS` file using the `readRDS` command.

```
autcon <- readRDS('PATH/autism_control.RDS')
```

Reveal the structure of the given dataset.

2. Perform the classification of the autism-control data using `rosetta`.

```
ros <- rosetta(autcon)
```

If you are not able to run R.ROSETTA you can use a saved model (`ros.RDS`).

- What is the quality of the model?
- How many rules did you obtain?

3. To display the rule network using the `visunetcyto()` function.

- Start Cytoscape Application
- Minimise the application
- In your R session, type

```
library(RCy3)
cytoscapePing ()
vis <- visunetcyto(ros$main)
```

The tool displays 10% of most significant nodes according to the connection value.

- Report the values of the rule filtration: minimum accuracy, minimum support, minimum decision coverage?
- For VisuNet networks in Cytoscape can you notice the separation between the networks? Use the `all` from drop-down menu.
- Can you identify the differences between two subnetworks, i.e., autism and control?
- Investigate the connections of nodes present on the networks. Find the strongest connected nodes for each decision.

To investigate a gene in autism netwok

- Find the hub gene, i.e, RHPN1.
- Right click on RHPN1 -> External Links -> Global Database Search -> NCBI
- Let us try another External Links search. Right click on RHPN1 -> External Links -> Publications -> Pubmed
- Report a potential link of RHPN1 with autism.

To save your network in Cytoscape.

- Go to File -> Export -> Network to Image...
- Export the networks of autism and control as a image of format .png.
- Browse to the folder where you wish to save the file and name the image as either autism.png or control.png

Assign the result to `vis`.

- Inspect the `vis` variable, what kind of the information is stored there?

4. The SFARI database stores the information about the autism-related genes. Check the overlap between genes present on the rule network and in the SFARI database. To obtain the gene names from the rule network use `vis$all$nodes$label`. File: `SFARI_Genes.RDS`.

- Find genes that overlapped both datasets.

```
SFARI <- readRDS('PATH/SFARI_Genes.RDS')

overlapped_genes <- SFARI[which(SFARI$gene.symbol %in% vis$all$nodes$label),]
```

What is the most significant gene for autism in reference to the gene score? More about the scoring procedure you can find here: https://gene.sfari.org/database/gene-scoring/

5. Investigate the GO terms associated with the VisuNet genes.

a. Install and load `clusterProfiler` and `org.Hs.eg.db` packages from `Bioconductor`.
b. Use function `bitr` to translate gene symbols to `entrezid`.

```
library(org.Hs.eg.db)
library(clusterProfiler)
gene_entrez <- bitr(unique(as.character( vis$all$nodes$label)), fromType = 'SYMBOL',
            toType = c("ENTREZID") , org.Hs.eg.db)
```

c. Apply function `groupGO` to obtain the GO annotations for the genes. Use the molecular functions `MF` as the ontology type, the GO level as 5 and `org.Hs.eg.db` as `OrgDb` parameter.

```
genes_GO <- groupGO(gene     = unique(gene_entrez$ENTREZID),
               OrgDb    = org.Hs.eg.db,
               ont      = "MF",
               level    = 5,
               readable = TRUE)
```

- What is the most frequent term? Find genes associated with this term.

Hint

Use a `summary` function to retrieve the information.

d. Remove terms that are associated with less than two gene. Create a `barplot` for the GO terms and the count of genes.

```
x1 <- summary(genes_GO)[which(summary(genes_GO)$Count>1),c('ID', 'Count')]
x <- as.numeric(x1[,2])
names(x)<- x1$ID
barplot(x, col = rainbow(20), las=2 )
```

- Try to interpret these terms in the context of autism.

Report

The report should include the R code, plots and answers. Also attach the networks from Cytoscape.