HW9

Tianqi Wu

4/10/2020

```
library(tidyverse)
library(biomaRt)
library(edgeR)
library(glmnet)
```

```
load("gtex_subset.RData") ## replace path with wherever you stored your file
expr = cpm(gtex_subset, log = TRUE)

expr = t(expr)
sum(apply(expr, 2, sd) == 0) ## no columns have zero variance
```

```
## [1] 0
```

```
set.seed(1) ## sets the random seed to get consistent cross-validation results
enet <- cv.glmnet(expr, tissue, alpha = 0.5, nfolds = 5, family = "binomial")
b <- coef(enet, s = "lambda.min")
```

Problem 1

1(a): BioMart

The gene symbols for the top 10 genes are presented in the R output. We could find all 10 matches but they are not in the right order.

```
elnet.coef = coef(enet, s = 'lambda.min')
elnet.coef@Dimnames[[1]][2:nrow(elnet.coef)] = gene_names

elnet.o = order(abs(elnet.coef), decreasing = TRUE)
top_10_gene = names(elnet.coef[elnet.o,][2:11])
top_10 = gsub("\\..*","",top_10_gene)

## top 10 genes
top_10
```

```
##  [1] "ENSG00000132670" "ENSG00000259203" "ENSG00000198382" "ENSG00000267675"
##  [5] "ENSG00000255399" "ENSG00000089225" "ENSG00000150977" "ENSG00000159423"
##  [9] "ENSG00000011454" "ENSG00000141505"
```

```
## biomart
## load mart
ensembl <- useMart("ensembl", dataset = "hsapiens_gene_ensembl")
```

```
## get gene_symbols
gene_symbols = getBM(attributes=c("ensembl_gene_id",
                  'external_gene_name'),
                  filters = 'ensembl_gene_id',
                  values = top_10, #name of the genes in a vector character file
                  mart = ensembl)

## gene symbols for top 10 genes
gene_symbols
```

```
##    ensembl_gene_id external_gene_name
## 1  ENSG00000011454            RABGAP1
## 2  ENSG00000089225               TBX5
## 3  ENSG00000132670              PTPRA
## 4  ENSG00000141505              ASGR1
## 5  ENSG00000150977             RILPL2
## 6  ENSG00000159423             ALDH4A1
## 7  ENSG00000198382               UVRAG
## 8  ENSG00000255399           TBX5-AS1
## 9  ENSG00000259203         AC016044.1
## 10 ENSG00000267675         AC105105.2
```

1(b): David

David mathces 300 gene ensemble ids out of 315 non zero ones and there are 56 cluseters. The top gene set is sequence variant with count 203, polymorphism with count 198, glycoprotein with count 137 and glycosylation site:N-linked (GlcNAc...) with count 130. The cluster with highest enrichment score is also presented.

Annotation Cluster 1	Enrichment Score: 19.6			Count	P_Value	Benjamini
UP_KEYWORDS	Secreted	RT		96	8.2E-30	2.4E-27
GOTERM_CC_DIRECT	extracellular region	RT		85	3.1E-25	8.1E-23
UP_SEQ_FEATURE	signal peptide	RT		118	5.0E-24	4.9E-21
UP_KEYWORDS	Glycoprotein	RT		137	4.4E-23	6.5E-21
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		130	1.7E-21	8.5E-19
UP_KEYWORDS	Signal	RT		125	2.1E-20	2.1E-18
GOTERM_CC_DIRECT	extracellular space	RT		61	1.7E-14	1.1E-12
UP_SEQ_FEATURE	disulfide bond	RT		87	2.0E-12	4.8E-10
UP_KEYWORDS	Disulfide bond	RT		94	6.5E-12	3.2E-10

Figure 1: David Cluster

```
## Get input for David
nonzero.gene = gene_names[b[-1] != 0]
ensembl_gene_id = gsub("\\..*","",nonzero.gene)
David.input = paste(ensembl_gene_id, collapse = "\n")
# cat(David.input)
```

Problem 2

1. Grand Challenge II-3: Develop genomebased approaches to prediction of disease susceptibility and drug response, early detection of illness, and molecular taxonomy of disease states.

2. Grand Challenge I-1: Comprehensively identify the structural and functional components encoded in the human genome.
3. Grand Challenge II-1: Develop robust strategies for identifying the genetic contributions to disease and drug response.

Problem 3

3.1 To Remember, the Brain Must Actively Forget

(https://www.quantamagazine.org/to-remember-the-brain-must-actively-forget-20180724/)

General Topic

The article focuses on how forgetting help remember better.

What. . . ?

What is mushroom body neurons that dopamine releases onto?

When. . . ?

When would the brain start to forget things?

Where. . . ?

Where are the forgotten memories stored?

How. . . ?

- How does protein Rac1 in the hippocampal neurons prolong the retention of memories?
- How did researchers increase the activity of Rac1?
- How does Rac1 involve in several forms of forgetting in fruit flies?

Why. . . ?

- Why would neurogenesis complicate the challenge of retrieving prior memories from the hippocampus?
- Why would added neural wiring damage the older engrams?
- Why would the overlaps make it harder to isolate the old memories from newer ones?

What if. . . ?

What if we decrease the activity of Rac1, would it let us forget some bad things happened recently?

I wonder if. . . ?

I wonder if there are some special SNPs that make Rac1 different from others.

Connection

This article reminds me of LSTM model in deep learning where we use forget gates to improve the performance.

3.2 Your Brain Chooses What to Let You See

(https://www.quantamagazine.org/your-brain-chooses-what-to-let-you-see-20190930/)

General Topic

The article focuses on how attention mechanism work by filtering unimportant things.

What...?

What are examples of automatic suppressive types of mechanisms?

When...?

When would automatic background subtraction take place?

Where...?

Where is the attention mechanism processed?

How...?

- How would we focus on important things by by lowering the priority of the rest?
- How does brain suppress information about the movement of the background?
- How can we perceive the movements of larger objects instead of smaller ones?

Why...?

- Why would older adults show little difference between perception of large and small objects?
- Why would they get much better at recognizing that motion with longer training period?
- Why would our brains adopt strategies that make smaller moving objects against those backgrounds stand out more?

What if...?

What if we do some experiments on larger predators that eat larger animals, would the result be the same?

I wonder if...?

I wonder if there are some specific neurons that make the attention strategy.

Connection

This article reminds me of attention model in deep learning where we only focus on the most relevant information

3.3 The Animal Origins of Coronavirus and Flu

(https://www.quantamagazine.org/how-do-animal-viruses-like-coronavirus-jump-species-20200225/)

General Topic

The article focuses on zoonoses, diseases that can jump between humans and other animals. In particular, it discusses the animal origins of coronavirus and influenza.

What. . . ?

What is the main difference between coronavirus and influenza?

When. . . ?

When does coronavirus result highest death rate?

Where. . . ?

Where are the receptors bound most by virus's proteins?

How. . . ?

- How does the virus reach the cells of the host?
- How does the virus recognize the cells of its host?
- How can S1 subunit of virus bind to the structure of the cells?

Why. . . ?

- Why is SARS-CoV-2 more deadly than SARS-CoV and MERS-CoV?
- Why bats and intermediate animals don't die from SARS-CoV-2?
- Why do people get infected by intermediate animals?

What if. . . ?

What if we mask the cells that are bound by the virus most, could the virus still recognize them?

I wonder if. . . ?

I wonder if there are some properties in cells that make bats and intermediate animals immune to SARS-CoV-2?

Connection

This article reminds me of GWAS analysis and we may use it to identify the similar virus that may be deadly to people and prepare for it.