Gene differential expression and disease prediction model

2017/6/28

Minstein

Data preparation:

Mice Protein Expression Data Set is downloaded from UCI's Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression).

The data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex. There are 38 control mice and 34 trisomic mice (Down syndrome), for a total of 72 mice. In the experiments, 15 measurements were registered of each protein per sample/mouse. Therefore, for control mice, there are 38x15, or 570 measurements, and for trisomic mice, there are 34x15, or 510 measurements. The dataset contains a total of 1080 measurements per protein. Each measurement can be considered as an independent sample/mouse.

Here we focus on genes differentially expressed between control mice and trisomic mice. To avoid influence by the injection treatment, we omit the measurements belonging to the memantine category and retain the saline group. Moreover, measurements including NA value are excluded and finally we obtain a dataset with 297 measurements on 77 protein/protein modifications (150 for control mice and 147 for trisomic mice).

Method and result:

We implement multiple t-test to determine the statistical significance of differential expression for each protein/protein modification. To deal with the negative positive problem, we use the Bonferroni correction to adjust the p-value. Other methods applicable for adjusting p-value including Benjamin-Hochberg (BH) and FDR, are also available in our program. The key commands for the test and computation are displayed below:

```
p_value_Genotype<-sapply(data1[,2:78],function(x)  t.test(x  ~
data1$Genotype)$p.value)
Bonferroni<-p.adjust(method="bonferroni",p_value_Genotype,n=77)
```

Given the p-value threshold as 0.001, we finally got 25 proteins/protein modifications, of which the differential expression is significant. They are listed by adjusted p-value in Table 1.

Table 1. The 25 differentially expressed genes in control mice and trisomic mice

Protein_or_modification	p_value_Genotype	Bonferroni
Tau_N	5.95E-43	4.58E-41
S6_N	5.99E-30	4.61E-28
pPKCG_N	7.12E-30	5.48E-28
APP_N	1.76E-28	1.36E-26
AcetylH3K9_N	1.31E-27	1.01E-25
ITSN1_N	6.01E-16	4.63E-14
DYRK1A_N	1.44E-13	1.10E-11
GluR3_N	6.48E-13	4.99E-11
H3AcK18_N	1.76E-12	1.36E-10
pNR1_N	1.94E-12	1.49E-10
P3525_N	2.02E-12	1.56E-10
pGSK3B_Tyr216_N	5.59E-12	4.30E-10
NR1_N	5.08E-09	3.91E-07
BAD_N	1.67E-08	1.28E-06
pNR2B_N	2.66E-08	2.05E-06
NR2B_N	3.64E-08	2.80E-06
MTOR_N	4.41E-08	3.39E-06
pP70S6_N	1.01E-07	7.76E-06
GFAP_N	2.71E-07	2.09E-05
AMPKA_N	3.52E-07	2.71E-05

P38_N	4.83E-07	3.72E-05
pMTOR_N	5.39E-07	4.15E-05
NR2A_N	7.11E-07	5.47E-05
BRAF_N	2.10E-06	0.000161436
pRSK_N	9.19E-06	0.000707883

Next, we use the above 25 proteins/protein modifications (total 297 measurements) to build a prediction model, so as to classify whether a mouse is trisomic or not. We applied 10-fold cross-validation to evaluate model's performance and perform the model selection and finally chosen LDA (Linear Discriminant Analysis) as our basic model. Our results in 10-fold cross-validation show no wrong predictions on the 297 mice with saline injection. As a control, we also applied the prediction model to the 255 mice with memantine injection (NA value excluded) and the wrong prediction ratio reaches 20%.

The calculated coefficients for the 25 proteins/protein modifications in our LDA model are listed in Table 2.

Table 2. The coefficients of 25 proteins/protein modifications in LDA model

Protein/Protein modification	LD1 coefficient	Protein/Protein modification	LD1 coefficient
DYRK1A_N	-1.10	NR2B_N	-5.01
ITSN1_N	11.19	pP70S6_N	8.82
NR1_N	-5.39	pPKCG_N	-0.66
NR2A_N	0.48	S6_N	-0.71
pNR1_N	2.21	AcetylH3K9_N	-11.55
pNR2B_N	0.00	Tau_N	30.91
pRSK_N	0.78	GFAP_N	18.66
BRAF_N	-8.68	GluR3_N	-22.00
APP_N	44.42	P3525_N	-7.86
MTOR_N	-8.20	pGSK3B_Tyr216_N	-1.11
P38_N	0.77	BAD_N	20.60
pMTOR_N	1.11	H3AcK18_N	-4.72
AMPKA_N	-23.83		

Code:

R language is used and all involved source code is included in gene_diff.R.