Random Gene Set Enrichment Analysis

Chengcheng Ma

October 27, 2020

1 Introduction

Random Gene Set Enrichment Analysis (RGSEA) is an algorithm for measuring the similarities between samples and classifies the samples based on transcriptome data. The algorithm combines bootstrap aggregating[1] and gene set enrichment analysis (GSEA)[2], which is similar to random forests[3] and random generalized linear model[4]. This algorithm is non-parametric and does not need to fit parameter, so the robustness of this method is high and there is no overfitting problem. Using this algorithm, researchers can compare the data from different studies or classify the samples with the data from other studies. There are three functions in this package: RGSEAfix, RGSEAsd and RGSEApredict.

RGSEAfix and RGSEAsd are the functions implementing RGSEA algorithm, RGSEApredict is for classification based on the results generated from RGSEA fix and RGSEAsd. Seven parameters are needed to be input for RGSEAfix.

1.1 Inputs

- query:a matrix, the query data
- reference: matrix, the reference data
- queryclasses: character vector, the classes of the query data
- refclasses: character vector, the classes of reference data
- random: numeric variable, the number of randomly sampled features to form the subset features.
- featurenum: numeric variable, the number of features selected from both top and bottom of the subset features
- iteration: numeric variable, the times of random sampling
- RGSEAsd includes all the seven parameters but
- instead a parameter named as sd, which indicates features with sd deviations from the mean value of the subset features be selected for the calculation.

RGSEAsd includes all the seven parameters but 5), instead a parameter named as sd, which indicates features with sd deviations from the mean value of the subset features be selected for the calculation. RGSEAprecdict is the function calculating the relative probability of the query data to be from a class of the refclasses based on the result generated by RGSEAfix or RGSEAsd.

2 Example Datset Description

For RGSEA, we need two datasets - The reference data is defined as the data, whose classes we already know. whereas query data is defined as the data, which we want to know the classes. In our given example, four samples from GDS4100, stored in (e2) is the reference data and two samples from GDS4102,stored in (e1) is the query data. The data were downloaded by command getGEO in GEOquery and transformed to expression datasets with GDS2eSet. The four samples from GDS4100 are GSM356796, GSM356797(tumor), GSM356828 and GSM356829(normal). The two samples from GDS4102 are GSM414924(tumor) and GSM414975(normal).

The cmap data is the part of the dataset of conncectivity map build 01 downloaded form

http://0-www.ncbi.nlm.nih.gov.elis.tmu.edu.tw/geo/query/acc.cgi?acc=GSE5258. An instance means the transcriptome of a cell line perturbed by a chemical and its corresponding negative control. We normalized the whole data, subtract the controls from the corresponding perturbations. The query ddata is 520276400579117512O104.C08, treated with thioridazine. The reference data are 520276400578914811290 4.G05, 520276400578914811290 4.F03,

520276400578914811290 4.F05, 520276400578914811290 4.E02, 520276400578914811290 4.E04. They were treated with tretinoin, prochlorperazine, chlorpromazine, vorinostat, sirolimus respectively. All the data were generated by MCF7 cell line.

2.1 Data download

We downloaded the file using :

```
library(GEOquery)
g4100 <- GDS2eSet(getGEO("GDS4100"))
g4102 <- GDS2eSet(getGEO("GDS4102"))
```

2.2 Data transformations

The data was then transformed in the following way

```
e4102<-exprs(g4102)
e4100<-exprs(g4100)
```

2.3 Final Exmple Data

```
e1<-e4102[,c(1,51)]
e2<-e4100[,c(1,2,23,24)]
colnames(e1)<-c("tumor", "normal")
colnames(e2)<-c("tumor","tumor","normal","normal")
```

This data was stored in variables e1 and e2 and is now availiable for you.

3 Running RGSEA

Here are two examples for how to use RGSEA Suppose we want to classify two samples from GDS4102 based on the data from GDS4100.

```
library(RGSEA)
data(e1)
data(e2)
RGSEAfix(e1,e2, queryclasses=colnames(e1), refclasses=colnames(e2), random=20000,
featurenum=1000, iteration=100)->test

## Processing #1...
## Processing #2...
```

We can see the result from "test".

```
test[[1]]

##        tumor tumor normal normal
## tumor     91     0      9      0
## normal    38     0     62      0
```

The column names are the class of reference data. The row names are the class of the query data. We can also predict the relative probability of the query data.

```
RGSEApredict(test[[1]], colnames(e2))

##        normal tumor
## tumor    0.09  0.91
## normal   0.62  0.38
```

Here, we can see that the relative probability of the probability to be each of the class.

Another example is measuring the similarities of the data from Connectivity map build 01.

The query data is treatment of MCF7 cell line by thioridazine. The reference data is treatment of MCF7 cell line by tretinoin, prochlorperazine, chlorpromazine, vorinostat and sirolimus respectivelly.

```
data(cmap)
test2<-RGSEAsd(cmap[,1],cmap[,2:6], queryclasses=colnames(cmap)[1],

refclasses=colnames(cmap)[2:6], random=5000, sd=2, iteration=100)

## Processing #1...

test2[[1]]

##      tretinoin prochlorperazine chlorpromazine vorinostat sirolimus
## [1,]         7               33             54          2         4
```

As we can see from the result, the value of chlorpromazine is the largest, which means among the five chemicals the function of chlorpromazine is most similar to thioridazine.

4 References

- Breiman L (1996) Bagging predictors. Machine learning 24: 123-140.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene setenrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America102: 15545-15550.
- Breiman L (2001) Random forests. Machine learning 45: 5-32.
- Song L, Langfelder P, Horvath S (2013) Random generalized linear model: a highly accurate and interpretable ensemble predictor. BMC bioinformatics 14: 5.
- Pei H, Li L, Fridley BL, Jenkins GD et al. FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. Cancer Cell 2009 Sep 8;16(3):259-66. PMID: 19732725
- Zhang L, Farrell JJ, Zhou H, Elashoff D et al. Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer. Gastroenterology 2010 Mar; 138(3):949-57.e1-7. PMID: 19931263