GENEMABR_PDF

Tao Fang, Daniel Marbach, Jitao David Zhang

4/28/2019

Contents

1 Introduction

GENEMABR is an R package for gene module annotation or gene set enrichment analysis via a regression based method. We formulate the gene set enrichment problem within a regression framework. Thus, the problem of gene-set enrichment is transformed into a feature selection problem, where the aim is to select the gene sets that best predict the membership of genes in a given gene set/module. Here we propose to apply regularised regression methods, such as lasso (l1 regularization), ridge (l2 regularization), or elastic net (hybrid of l1 and l2 regularization controlled by the hyperparameter alpha), in order to adjust the treatment of similar or redundant gene sets. For more details about this method. Please refer to out paper:(?)

2 Installation

GENEMABR can be installed from Bioconductor:

```
if (!requireNamespace("BiocManager", quietly=TRUE)){
    install.packages("BiocManager")}
BiocManager::install("GENEMABR")
```

Altertivly, you can install GENEMABAR via git devltools?

The package can be loaded using the `library` command.

```
library(GENEMABR)

## load other required package to run GENEMABR
if(!require(glmnet)){
    install.packages("glmnet")
    library(glmnet)
}
if(!require(Matrix)){
    install.packages("Matrix")
    library(Matrix)
}
if(!require(igraph)){
    install.packages("igraph")
    library(igraph)
}
```

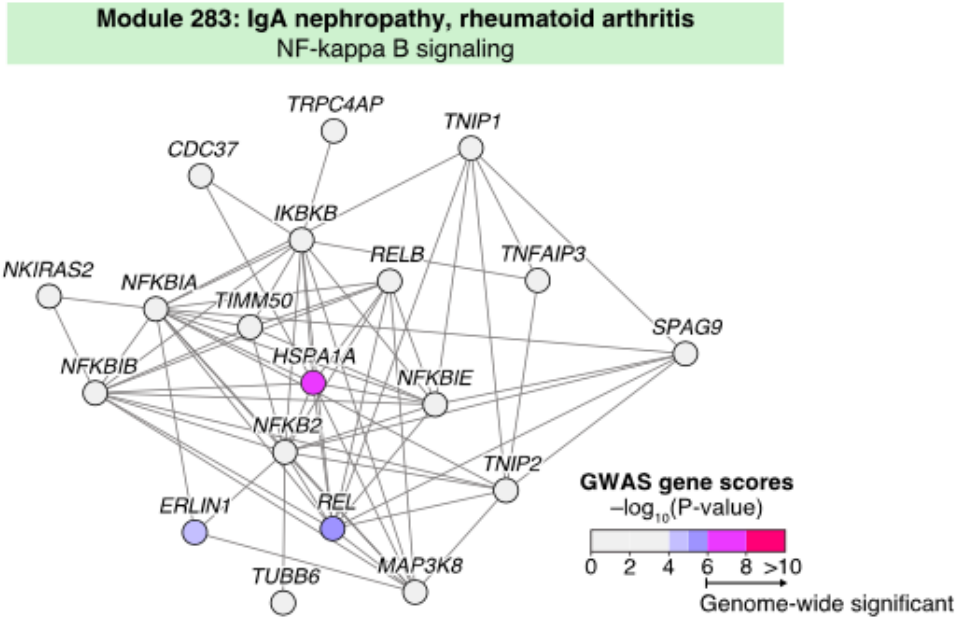
To see the latest updates and releases or to post a bug, see our GitHub page at https://github.com/TaoDFang/GENEMABR. To ask questions about running GENEMABR, plase create issures at https://github.com/TaoDFang/GENEMABR/issues

3 Gene module annotation or gene set enrichment analysis via regression based method

To test a our method, we use a IgAN-associated module from the paper: Choobdar,S. et al. (2019) Open Community Challenge Reveals Molecular Network Modules with Key Roles in Diseases. bioRxiv, 265553.

IgAN-associated module identified using the consensus analysis in the InWeb protein-protein interaction network. The module comprises immune-related NF-B signaling pathways.

Figure below shows mannually extracted P-values for some important pathways from GO Ontology and REACTOME database by using the non-central hypergeometric distribution test.



GO biological process	P-value
i-kappab kinase/nf-kappab signaling	1.4E-10
regulation of innate immune response	2.27E-08
positive regulation of nf-kappab transcription factor activity	3.44E-08
innate immune response-activating signal transduction	1.41E-06
stress-activated mapk cascade	1.44E-06
stress-activated protein kinase signaling cascade	1.62E-06
activation of innate immune response	1.65E-06
positive regulation of sequence-specific dna binding tf activity	1.73E-06
regulation of tumor necrosis factor-mediated signaling	2.92E-06
pattern recognition receptor signaling pathway	3.83E-06

Reactome pathways	P-value
rip mediated nfkb activation via dai	3.03E-08
traf6 mediated nfkb activation	3.94E-08
tak1 activates nfkb by phosphorylation and activation of ikks	7.65E-08
rig i mda5 mediated induction of ifn alpha beta pathways	1.4E-07
activation of nf kappab in b cells	6.95E-07
downstream signaling events of b cell receptor bcr	7.49E-06
il1 signaling	1.33E-05
signaling by the b cell receptor bcr	2.84E-05
nfkb and map kinases activation mediated by tlr4 signaling	3.67E-05
traf6 mediated induction of nfkb and map kinases upon tlr7 8 or 9 activation	3.99E-05

Read non-central hypergeometric distribution test results from original paper for gene module based on GO Ontology and REACTOME pathways

```
hypergeometric_test_results=read.csv(file = "../data_raw/Daniel_S5_mod283.txt",header = T,sep = "\t")
## filterd results with only P.noncentral value less then 0.05
hypergeometric_test_results_filtered=hypergeometric_test_results[hypergeometric_test_results$P.noncentral<0.0

head(hypergeometric_test_results_filtered)
##   pathwayDb   network                    module
## 1     go_bp PPI-InWeb PPI-InWeb_Consensus_mod283
## 2     go_bp PPI-InWeb PPI-InWeb_Consensus_mod283
## 3     go_bp PPI-InWeb PPI-InWeb_Consensus_mod283
## 4     go_bp PPI-InWeb PPI-InWeb_Consensus_mod283
## 5     go_bp PPI-InWeb PPI-InWeb_Consensus_mod283
## 6     go_bp PPI-InWeb PPI-InWeb_Consensus_mod283
##                                                    term     termId
## 1               glutathione derivative biosynthetic process GO:1901687
## 2                          cell cycle G2/M phase transition GO:0044839
## 3            immune response-activating signal transduction GO:0002757
## 4     innate immune response-activating signal transduction GO:0002758
## 5 proteolysis involved in cellular protein catabolic process GO:0051603
## 6                           stress-activated MAPK cascade GO:0051403
##   P.hypergeom P.noncentral P.hypergeom.bonf P.noncentral.bonf
## 1    0.050400     0.047000                1                 1
## 2    0.040700     0.024100                1                 1
## 3    0.000568     0.000378                1                 1
## 4    0.000227     0.000199                1                 1
## 5    0.030900     0.021400                1                 1
## 6    0.030800     0.025100                1                 1
##   P.hypergeom.fdr P.noncentral.fdr
## 1           1.000           1.0000
## 2           1.000           1.0000
## 3           0.131           0.1240
## 4           0.063           0.0743
## 5           1.000           1.0000
## 6           1.000           1.0000
dim(hypergeometric_test_results_filtered)
## [1] 176  11
```

We found traditional non-central hypergeometric distribution test/fisher extact test usually return too much over-presented pathways for certain gene module.

And many of these pathways are highly related. For example,four pathways with pvalues less than 0.05:

immune response-activating cell surface receptor signaling pathway GO:0002429;

immune response-activating signal transduction GO:0002757 ;

immune response-regulating cell surface receptor signaling pathway GO:0002768;

immune response-regulating signaling pathway: GO:0002764.

GO:0002757 and GO:0002768 are parents of GO:0002429 and GO:0002764 is parents of GO:0002757 and GO:0002768

To get a more sparse results by exploting corraltion between different pathways, we could use "regression_selected_pathways" function in GENEMABR package.

This function use regularised regression methods to do enrichment analysis, such as lasso (l1 regularization), ridge (l2 regularization), or elastic net (hybrid of l1 and l2 regularization controlled by the hyperparameter alpha), in order to adjust the treatment of similar or redundant gene-sets. If two gene-sets are highly redundant, lasso will assign a higher coefficient to one of them randomly, ridge will assign equal coefficients to both of them, whereas elastic net will behave between lasso and ridge.

To use this method, use need to provide a binary gene pathways realationships matrix whose columns are the pathways/gene sets and whose rows are all the genes from pathways/gene sets.For gene i and pathway j, the value of matrix(i,j) is 1 is gene i belonging to pathway j otherwise 0.

Users could use default gene_pathway_matrix("default") so it will use pre-collected gene_pathway_matrix from GO Ontology and REACTOME databaase.Altertatively, Users could use their own customized gene_pathway_matrix

```
#Gene module from the paper
gene_list=c("TRPC4AP","CDC37","TNIP1","IKBKB","NKIRAS2","NFKBIA","TIMM50","RELB","TNFAIP3","NFKBIB","HSPA1A",
#help("regression_selected_pathways")

#Here use regression_selected_pathways with default gene pathway matrix and set the alpha value as 0.5
enrichment_results=regression_selected_pathways(gene_input=gene_list,gene_pathway_matrix="default",alpha=0.5)
enrichment_results
## $selected_pathways_names
## $selected_pathways_names$`R-HSA-1810476`
## [1] "RIP-mediated NFkB activation via ZBP1"
##
## $selected_pathways_names$`R-HSA-5603029`
## [1] "IkBA variant leads to EDA-ID"
##
## $selected_pathways_names$`GO:0032688`
## [1] "negative regulation of interferon-beta production"
##
## $selected_pathways_names$`R-HSA-933542`
## [1] "TRAF6 mediated NF-kB activation"
##
## $selected_pathways_names$`GO:0007249`
## [1] "I-kappaB kinase/NF-kappaB signaling"
##
## $selected_pathways_names$`R-HSA-1606322`
## [1] "ZBP1(DAI) mediated induction of type I IFNs"
##
##
## $selected_pathways_coef
## R-HSA-1810476 R-HSA-5603029    GO:0032688  R-HSA-933542    GO:0007249
##    0.12593491    0.07105989    0.01993546    0.01910265    0.00513055
## R-HSA-1606322
##    0.00219621
##
## $selected_pathways_fisher_pvalue
```

```
## R-HSA-1810476 R-HSA-5603029    GO:0032688  R-HSA-933542    GO:0007249
##  3.511133e-10  4.005332e-08  6.338554e-05  1.929685e-09  4.985802e-11
## R-HSA-1606322
##  7.596824e-10
##
## $selected_pathways_num_genes
## R-HSA-1810476 R-HSA-5603029    GO:0032688  R-HSA-933542    GO:0007249
##            11             7            11            16            63
## R-HSA-1606322
##            13
```

From results above, we can find our methods give a much more sparse while biological meaning results.It captures most important NF-kappaB signaling pathways from the gene module

4 Other functions in this methods

```
## if you use the default pathway databases(GO Ontologyand REACTOME).
## After you extracted enriched pathways, you can use find_root_ids function to find thier GO sub-root or REA
## Here we use GO sub-root instead of GO root nodes as there are only three roots in the GO ontology and it's
GO_Reactome_root_id=find_root_ids(selected_pathways=names(enrichment_results$selected_pathways_coef))
GO_Reactome_root_id
## $`R-HSA-1810476`
## [1] "R-HSA-168256"
##
## $`R-HSA-5603029`
## [1] "R-HSA-1643685"
##
## $`GO:0032688`
## [1] "GO:0065007"
##
## $`R-HSA-933542`
## [1] "R-HSA-168256"
##
## $`GO:0007249`
## [1] "GO:0009987#GO:0065007"
##
## $`R-HSA-1606322`
## [1] "R-HSA-168256"

# Or if you want to obatain root notes names instead of ID, you can use function from_id2name to get names fr
GO_Reactome_root_id_names=from_id2name(GO_Reactome_root_id)
GO_Reactome_root_id_names
## $`R-HSA-168256`
## [1] "Immune System"
##
## $`R-HSA-1643685`
## [1] "Disease"
##
```

```
## $`GO:0065007`
## [1] "biological regulation"
##
## $`R-HSA-168256`
## [1] "Immune System"
##
## $`GO:0009987#GO:0065007`
## [1] "cellular process"      "biological regulation"
##
## $`R-HSA-168256`
## [1] "Immune System"

# Or you  can use function get_steps function to calculate the distance from selected pathways to GO or React
step2root=get_steps(selected_pathways=names(enrichment_results$selected_pathways_coef))
step2root
## $`R-HSA-1810476`
## [1] 4
##
## $`R-HSA-5603029`
## [1] 3
##
## $`GO:0032688`
## [1] 7
##
## $`R-HSA-933542`
## [1] 3
##
## $`GO:0007249`
## [1] 4
##
## $`R-HSA-1606322`
## [1] 3
# To view specic position of GO/REACOTEM pathways in ontology trees.
# You can use Visualization tool at https://www.ebi.ac.uk/QuickGO/ and https://reactome.org/PathwayBrowser/
```

5 Session Information

```
sessionInfo()
## R version 3.6.0 (2019-04-26)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
```

```
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] igraph_1.2.4.1   glmnet_2.0-16    foreach_1.4.4    Matrix_1.2-17
## [5] GENEMABR_0.99.0  BiocStyle_2.11.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.1         bookdown_0.9       codetools_0.2-16
##  [4] lattice_0.20-38    digest_0.6.18      grid_3.6.0
##  [7] magrittr_1.5       evaluate_0.13      stringi_1.4.3
## [10] rmarkdown_1.12     iterators_1.0.10   tools_3.6.0
## [13] stringr_1.4.0      xfun_0.6           yaml_2.2.0
## [16] compiler_3.6.0     pkgconfig_2.0.2    BiocManager_1.30.4
## [19] htmltools_0.3.6    knitr_1.22
```