PALMER: A Constrained Biclustering Algorithm to Improve Pathway Annotation Based on the Biomedical Literature Mining with ‘palmer’ Package

Jin Hyun Nam[1] , Daniel couch[1] , Willian A. da Silveira[2,3] , Zhenning Yu[1] , and Dongjun Chung[1]
[1] Department of Public health Sciences, Medical University of South Carolina, Charleston, SC, USA. [2] Department of Pathology and Laboratory Medicine, Medical University of South Carolina, Charleston, SC, USA. [3] Center for Genomic Medicine, Medical University of South Carolina, Charleston, SC, USA.

April 12, 2019

1. Overview

This vignette provides an introduction to the genetic analysis using the ‘**palmer**’ package. R package ‘**palmer**’ implements PALMER (constrained biclustering algorithm to improve **P**athway **A**nnotation based on the **B**iomedical **L**iterature miming), a constrained biclustering approach that allows to identify indirect relationships among genes based on the text mining of biomedical literature, which allows researchers to utilize prior biological knowledge to guide identification of gene-gene associations.

The package can be loaded with the command:

```
library("palmer")
```

Users can find the most up-to-date versions of ‘**palmer**’ package in our GitHub webpage https://dongjunchung.github.io/palmer/

2. Exploration of Data

In this vignette, we use a example data set which are composed of binary matrix to be clusterd and two gene sets (pathways) as constraints. The binary matrix is size of 100 genes $\times$ 100 GO terms matrix, 1 value indicates that a gene is associated with a GO terms and 0 otherwise. First pathways is set of 27 genes and second one is set of 23 genes, which are randomly selected from the 100 gene list of binary matrix. Thus, assumption of data is that both 27 genes and 23 genes are known as gene pathwy and remaining 50 genes are clusterd to each pathway.

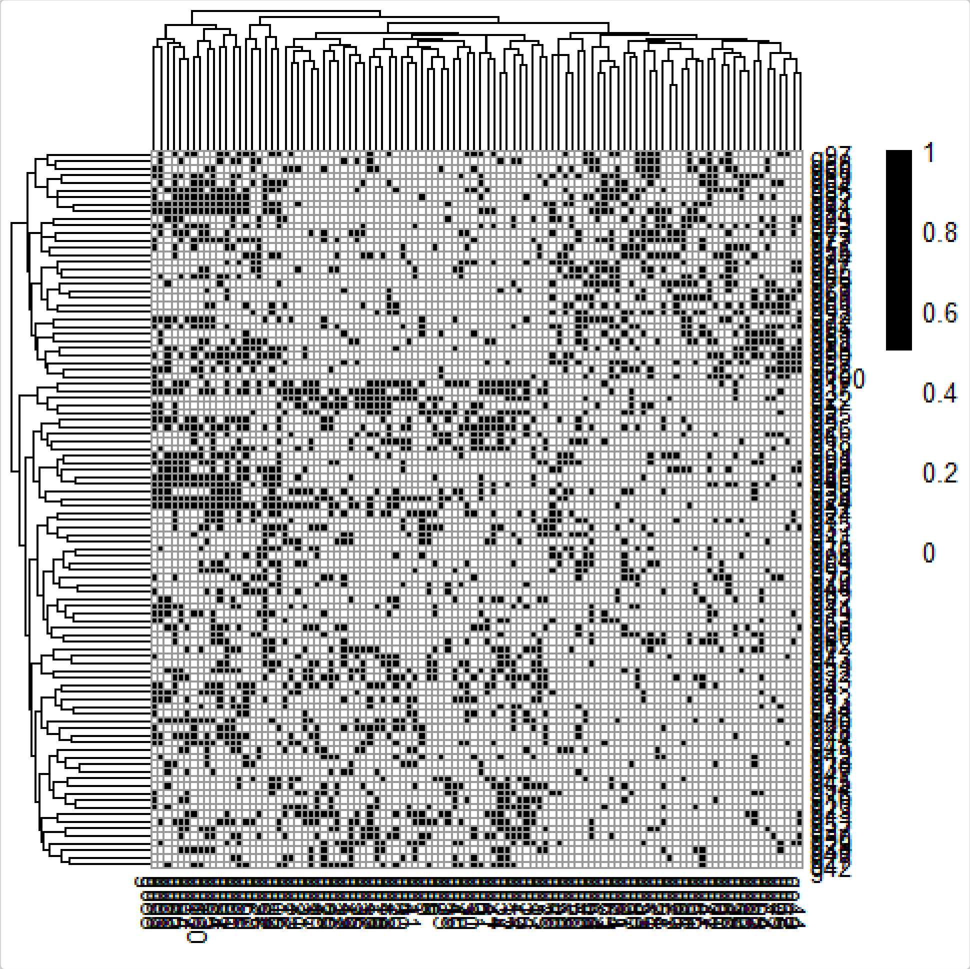
The data sets can be loaded with the command:

```
data(sdata)
data(pathway)
```

```
dim(sdata)
#> [1] 100 100
sdata[1:6,1:6]
#>    go1 go2 go3 go4 go5 go6
#> g1   0   1   1   1   0   1
#> g2   1   0   0   0   0   0
#> g3   1   1   1   0   1   0
#> g4   1   0   1   1   0   1
#> g5   0   1   0   0   0   0
#> g6   1   0   0   0   0   1
```

```
pathway
#> [[1]]
#>  [1] "g13" "g52" "g27" "g49" "g53" "g51" "g5"  "g19" "g33" "g6"  "g15"
#> [12] "g31" "g17" "g1"  "g3"  "g38" "g18" "g14" "g16" "g47" "g41" "g10"
#> [23] "g36" "g21" "g4"  "g30" "g2"
#>
#> [[2]]
#>  [1] "g68" "g75" "g99" "g95" "g66" "g80" "g96" "g64" "g85" "g61" "g73"
#> [12] "g88" "g97" "g76" "g83" "g71" "g92" "g86" "g78" "g90" "g93" "g74"
#> [23] "g65"
```

```
library(pheatmap)
#> Warning: package 'pheatmap' was built under R version 3.5.2
pheatmap(sdata,color=c("white","black"))
```



3. Fitting the PALMER

We are now ready to fit a PALMER using binary matrix and pathway information described above (sdata and pathway). PAMLER is essentially a biclustering algorithm, where conditional mixture models for genes and GO terms are fitted iteratively and confidence of these identifications is estimated using a blockspecific bootstrap procedure. R package palmer provides convience function ‘palmer’ for fitting PALMER and just need three arguments except for dataset and pathway, which are number of gene cluster, number of GO term cluster and number of bootstrapping for estimation of uncertainy of gene clustering. In this vignette, we set 2 for gene cluster number (K=2) under data assumption and set 3 for GO term cluster number (L=3). For the bootstraps, we set 100 (B=100, default is 1000).

We fit the PALMER with the command:

```
fit.palmer <- palmer(X=sdata,path=pathway,K=2,L=3,B=100)
```

The following command prints out a summary of PALMER fit, including data summary and model setting.

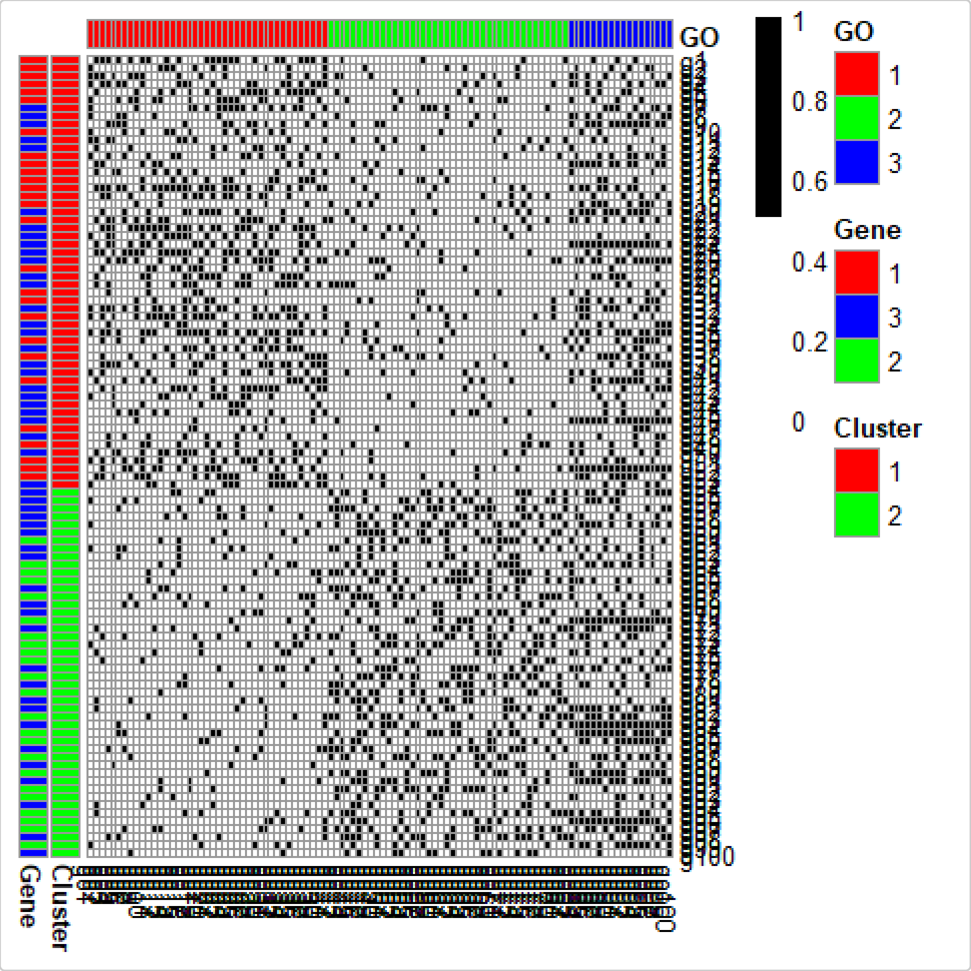
```
fit.palmer
#> Summary: palmer (class: palmer)
#> --------------------------------------------------
#> Model settings
#> Number of samples to be analyzed: 100
#> Number of variables to be analyzed: 100
#> Number of gene clusters: 2
#> Number of GO term clusters: 3
#> Number of bootstrapping: 100
#> --------------------------------------------------
```

Gene cluster and GO term cluster information and probability corresponding to each cluster can be extracted using method `predict’.

```
predict(fit.palmer)
#> $Gene
#>      True gene cluster Predicted gene cluster Probability
#> g1                   1                      1        1.00
#> g2                   1                      1        1.00
#> g3                   1                      1        1.00
#> g4                   1                      1        1.00
#> g5                   1                      1        1.00
#> g6                   1                      1        1.00
#> g7                   3                      1        1.00
#> g8                   3                      1        1.00
#> g9                   3                      1        1.00
#> g10                  1                      1        1.00
#> g11                  3                      1        1.00
#> g12                  3                      1        1.00
#> g13                  1                      1        1.00
#> g14                  1                      1        1.00
#> g15                  1                      1        1.00
#> g16                  1                      1        1.00
#> g17                  1                      1        1.00
#> g18                  1                      1        1.00
#> g19                  1                      1        1.00
#> g20                  3                      1        1.00
#> g21                  1                      1        1.00
#> g22                  3                      1        1.00
#> g23                  3                      1        1.00
#> g24                  3                      1        1.00
#> g25                  3                      1        1.00
#> g26                  3                      1        1.00
#> g27                  1                      1        1.00
#> g28                  3                      1        1.00
#> g29                  3                      1        1.00
#> g30                  1                      1        1.00
#> g31                  1                      1        1.00
#> g32                  3                      1        1.00
#> g33                  1                      1        1.00
#> g34                  3                      1        1.00
#> g35                  3                      1        1.00
#> g36                  1                      1        1.00
#> g37                  3                      1        1.00
#> g38                  1                      1        1.00
#> g39                  3                      1        1.00
#> g40                  3                      1        1.00
#> g41                  1                      1        1.00
#> g42                  3                      1        1.00
#> g43                  3                      1        1.00
#> g44                  3                      1        1.00
#> g45                  3                      1        1.00
#> g46                  3                      1        1.00
#> g47                  1                      1        1.00
#> g48                  3                      1        1.00
#> g49                  1                      1        1.00
#> g50                  3                      1        1.00
#> g51                  1                      1        0.99
#> g52                  1                      1        0.72
#> g53                  1                      1        0.98
#> g54                  3                      1        0.88
#> g55                  3                      2        0.96
#> g56                  3                      2        0.95
#> g57                  3                      2        1.00
#> g58                  3                      2        1.00
#> g59                  3                      2        1.00
#> g60                  3                      2        1.00
#> g61                  2                      2        1.00
#> g62                  3                      2        1.00
#> g63                  3                      2        1.00
#> g64                  2                      2        1.00
#> g65                  2                      2        0.93
#> g66                  2                      2        1.00
#> g67                  3                      2        1.00
#> g68                  2                      2        1.00
#> g69                  3                      2        1.00
#> g70                  3                      2        1.00
#> g71                  2                      2        1.00
#> g72                  3                      2        0.92
#> g73                  2                      2        1.00
#> g74                  2                      2        0.70
#> g75                  2                      2        1.00
#> g76                  2                      2        1.00
#> g77                  3                      2        1.00
#> g78                  2                      2        0.99
#> g79                  3                      2        1.00
#> g80                  2                      2        1.00
#> g81                  3                      2        0.98
#> g82                  3                      2        1.00
#> g83                  2                      2        1.00
#> g84                  3                      2        0.96
#> g85                  2                      2        0.82
#> g86                  2                      2        0.98
#> g87                  3                      2        0.79
#> g88                  2                      2        0.96
#> g89                  3                      2        1.00
#> g90                  2                      2        0.99
#> g91                  3                      2        0.99
#> g92                  2                      2        1.00
#> g93                  2                      2        0.99
#> g94                  3                      2        1.00
#> g95                  2                      2        1.00
#> g96                  2                      2        0.89
#> g97                  2                      2        0.98
#> g98                  3                      2        0.96
#> g99                  2                      2        1.00
#> g100                 3                      2        1.00
#> 
#> $GO
#>       GO term cluster Probability
#> go1                 1   0.9999889
#> go2                 1   0.9999375
#> go3                 1   0.9999999
#> go4                 1   0.9999999
#> go5                 1   1.0000000
#> go6                 1   0.9997633
#> go7                 1   0.9996265
#> go8                 1   0.9997969
#> go9                 1   1.0000000
#> go10                1   0.9999999
#> go11                1   0.9999982
#> go12                1   0.9999982
#> go13                1   0.9994925
#> go14                1   0.9999999
#> go15                1   0.9999984
#> go16                1   0.9972166
#> go17                1   0.9998371
#> go18                1   0.9999995
#> go19                1   0.9999999
#> go20                1   0.9999120
#> go21                1   0.9999977
#> go22                1   0.9999806
#> go23                1   0.9999384
#> go24                1   0.9999938
#> go25                1   0.9999897
#> go26                1   0.9999994
#> go27                1   0.9999299
#> go28                1   1.0000000
#> go29                1   0.9999977
#> go30                1   0.9996089
#> go31                1   0.9994925
#> go32                1   0.9999995
#> go33                1   0.9999900
#> go34                1   0.9995579
#> go35                1   0.9999973
#> go36                1   0.9999897
#> go37                1   0.9999959
#> go38                1   0.9994925
#> go39                1   0.9999375
#> go40                1   0.9999384
#> go41                1   0.9998124
#> go42                2   0.9999998
#> go43                2   0.9994596
#> go44                2   0.9999613
#> go45                2   0.9998880
#> go46                2   0.9997926
#> go47                2   0.9999992
#> go48                2   0.9999718
#> go49                2   0.9579671
#> go50                2   0.9888526
#> go51                2   0.9968148
#> go52                2   0.9999889
#> go53                2   0.9999952
#> go54                2   0.9999990
#> go55                2   0.9999997
#> go56                2   0.9999872
#> go57                2   0.9991912
#> go58                2   0.9994416
#> go59                2   0.9999999
#> go60                2   0.9999999
#> go61                2   0.9988900
#> go62                2   0.9999613
#> go63                2   0.9999974
#> go64                2   1.0000000
#> go65                2   0.9999790
#> go66                2   0.9999990
#> go67                2   0.9888526
#> go68                2   0.9997633
#> go69                2   1.0000000
#> go70                2   0.9996074
#> go71                2   0.9999947
#> go72                2   0.9999718
#> go73                2   0.9999961
#> go74                2   0.9999965
#> go75                2   0.9960290
#> go76                2   0.9984735
#> go77                2   1.0000000
#> go78                2   0.9999999
#> go79                2   0.9999718
#> go80                2   0.9999718
#> go81                2   0.9999997
#> go82                2   0.9999994
#> go83                3   0.9999999
#> go84                3   0.9999870
#> go85                3   0.9999470
#> go86                3   0.9999998
#> go87                3   0.9992959
#> go88                3   1.0000000
#> go89                3   0.9996109
#> go90                3   0.9993146
#> go91                3   1.0000000
#> go92                3   0.9999982
#> go93                3   0.9999997
#> go94                3   0.8999412
#> go95                3   1.0000000
#> go96                3   0.9994131
#> go97                3   0.9998657
#> go98                3   0.9999937
#> go99                  3    0.9999725
#> go100                 3    0.9998970
```

A method ‘plot’ shows heatmap with cluster information.

```
plot(fit.palmer)
```



4. Web Interface: LitSelect

In order to further facilitate users’ convenience to access and obtain the literature mining data for PALMER analysis, we developed the web interface LitSelect, LitSelect uses GAIL’s (http://chunglab.io/GAIL) graph database to access low-level literature mining data (p-values) and then perform candidate gene and GO term selection. LitSelect is accessible at http://chunglab.io/LitSelect. The interface contains a form with the following input fields for the user to provide: 1) two gene lists, 2.) a ratio of candidate genes to select, 3.) a ratio of candidate GO terms to select, and 4.) a binarization cutoff. To ensure unique mapping of gene IDs, the interface requires HGNC IDs. The interface contains a link to an ID mapper we previously developed in case the user has some other name for the gene (e.g. gene symbol). LitSelect then performs the selection of candidate gene and GO terms based on the user-provided parameters and pathway information and then we can download the literature mining data for PALMER.