How To Use PAGI

Junwei Han,Yanjun Xu,Haixiu Yang,Chunquan Li and Xia Li

December 12, 2012

Contents

1 Overview

This vignette demonstrates how to easily use the PAGI package. This package can identify canonical KEGG pathways associated with two different biological states. Our system provides a new strategies of identifying pathways based on global influnce based on global influence from both the internal effect of pathways and crosstalk between pathways(see the section 2).

2 Identifying canonical biological pathways based on global influence from both the internal effect of pathways and crosstalk between pathways

The section introduces our pathway analysis based on global influence (PAGI) method for identifying canonical biological pathways associated with different biological states. PAGI used a network-based approach to find the latent dysregulated pathways by considering the global influence from both the internal effect of pathways and crosstalk between pathways. Firstly, we constructed a global gene-gene network based on the relationships of genes extracted from each pathway in KEGG database and the overlapped genes between pathways. The global gene-gene network data is stored in the environmental variable (netWorkdata). The expression profiles data with normal and disease samples were mapped to the global network. Then we defined a global influence factor (GIF) to distinguish the non-equivalence of gene influenced by both internal effect of pathways and crosstalk between pathways in the global network. The random walk with restart (RWR) algorithm was used to evaluate the GIF score by integrating the global network topology and the correlation of gene with phenotype (see the section 2.1). We used the function `CalGIF` to calculate the GIF scores. Finally, we used cumulative distribution functions (CDFs) to prioritize the dysregulated pathways ((see the section 2.2)). We used the function `PAGI.Main` to prioritize the pathways.

2.1 Calculating the scores of global influence factors (GIFs)

The random walk with restart (RWR) algorithm was used to evaluate the GIF by integrating the global network topology and the correlation of gene with phenotype.

The function `CalGIF` can calculate the GIF scores of genes in the gene expression data which is inputed by user. The following commands can calculate the scores of GIFs in a given dataset.

```
> #example 1
> #get example data
> dataset<-getdataset()
> class.labels<-getclass.labels()
> #calculate the global influence factor (GIF)
> GIFscore<-CalGIF(dataset,class.labels)
> #print the top ten results to screen
> GIFscore[rev(order(GIFscore))][1:10]

     TP53    CDKN1A       BAX      GNAL      GNAS      MDM2     ACTG1    MAPK11
1.0000000 0.9669012 0.9070069 0.8073090 0.7538116 0.7172111 0.7023068 0.6974868
    STAT6      DDB2
0.6461288 0.6283938

> #example 2
> #get example data
> dataset<-read.table(paste(system.file(package="PAGI"),"/localdata/dataset.txt",sep=""),
+ header=T,sep="\t",quote="\"")
> class.labels<-as.character(read.table(paste(system.file(package="PAGI"),
+ "/localdata/class.labels.txt",sep=""),quote="\"", stringsAsFactors=FALSE)[1,])
> #calculate the global influence factor (GIF)
> GIFscore<-CalGIF(dataset,class.labels)
> #print the top ten results to screen
> GIFscore[rev(order(GIFscore))][1:10]

     TP53    CDKN1A       BAX      GNAL      GNAS      MDM2     ACTG1    MAPK11
1.0000000 0.9669012 0.9070069 0.8073090 0.7538116 0.7172111 0.7023068 0.6974868
    STAT6      DDB2
0.6461288 0.6283938
```

2.2 Identifying pathways based on global influence

The function `PAGI.Main` can identify dysregulated pathways which may be associated with two bilogical states. The result is a list. It includes two elements: summary result and pathway list. Summary result is a dataframe. It is the summary of the result of pathways. Each rows of the dataframe represents a pathway. Its columns include "Pathway Name", "SIZE", "PathwayID", "Pathway Score", "NOM p-val", "FDR q-val", "Tag percentage" (Percent of gene set before running enrichment peak), "Gene percentage" (Percent of gene list before running enrichment peak), "Signal strength" (enrichment signal strength). Pathway list is of pathways which present the detail results of pathways with NOM p-val< p.val.threshold or FDR< FDR.threshold. Each element of the list is a dataframe. Each rows of the dataframe represents a gene. Its columns include "Gene number in the (sorted) pathway", "gene symbol from the gene express data", "location of the gene in the sorted gene list", "the T-score of gene between two biological states", "global influence impactor", "if the gene contribute to the score of pathway". The following commands can identify the dysregulated pathways in a given dataset with default parameters.

```
> #example 1
> #get example data
> dataset<-getdataset()
> class.labels<-getclass.labels()
> #identify dysregulated pathways
> result<-PAGI.Main(dataset,class.labels,nperm = 100,p.val.threshold = -1,FDR.threshold = 0.01,
+ gs.size.threshold.min = 25, gs.size.threshold.max = 500 )

[1] "Running PAGI Analysis..."

> #print the summary results of top ten pathways to screen
> result[[1]][1:10,]

                          Pathway Name SIZE     PathwayID Pathway Score
1                ErbB signaling pathway   73 path:hsa04012       0.53419
2             Calcium signaling pathway  146 path:hsa04020       0.49004
3                           Cell cycle   99 path:hsa04110       0.48449
4                       Oocyte meiosis   76 path:hsa04114       0.48058
5                 p53 signaling pathway   49 path:hsa04115       0.73209
6                           Apoptosis   73 path:hsa04210       0.52263
7                VEGF signaling pathway   56 path:hsa04370       0.49866
8        Cell adhesion molecules (CAMs)   95 path:hsa04514       0.44334
9                         Gap junction   63 path:hsa04540       0.48412
10 Toll-like receptor signaling pathway   83 path:hsa04620       0.49664
   NOM p-val FDR q-val Tag \\% Gene \\% Signal
1          0         0   0.151   0.0658  0.142
2          0         0    0.26    0.145  0.226
3          0         0   0.333    0.273  0.245
4          0         0   0.434    0.286  0.312
5          0         0   0.184   0.0358  0.178
6          0         0   0.164   0.0688  0.154
7          0         0   0.232   0.0782  0.215
8          0         0   0.337    0.234  0.261
9          0         0   0.286    0.138  0.248
10         0         0   0.301    0.172  0.251

> #print the detail results of top ten genes in the first pathway to screen
> result[[2]][[1]][1:10,]

    # GENE SYMBOL LIST LOC   Tscore(p-value)   GIF CORE_ENRICHMENT
1   1      CDKN1A        1 5.96 ( 1.44e-07 ) 0.967             YES
2   2         SRC       49  2.53 ( 0.00737 ) 0.512             YES
3   3      CAMK2A       75  2.49 ( 0.00814 ) 0.421             YES
4   4      MAP2K1       82  2.45 ( 0.00899 ) 0.419             YES
5   5      PIK3CA      118   2.08 ( 0.0214 ) 0.606             YES
6   6      MAP2K7      132   2.16 ( 0.0179 )   0.5             YES
7   7      CAMK2B      189   2.18 ( 0.0171 ) 0.375             YES
8   8       PLCG2      246   1.96 ( 0.0279 ) 0.493             YES
9   9        NRAS      461   1.82 ( 0.0375 ) 0.432             YES
10 10        PAK3      644    1.8 ( 0.0391 )  0.32             YES

> #write the summary results of pathways to tab delimited file.
> write.table(result[[1]], file = "SUMMARY RESULTS.txt", quote=F, row.names=F, sep = "\t")
```

```
> #write the detail results of genes for each pathway with FDR.threshold< 0.01 to tab delimited file.
> for(i in 1:length(result[[2]])){
+ gene.report<-result[[2]][[i]]
+ filename <- paste(names(result[[2]][i]),".txt", sep="", collapse="")
+ write.table(gene.report, file = filename, quote=F, row.names=F, sep = "\t")
+ }
> #example 2
> #get example data
> dataset<-read.table(paste(system.file(package="PAGI"),"/localdata/dataset.txt",sep=""),
+ header=T,sep="\t",quote="\"")
> class.labels<-as.character(read.table(paste(system.file(package="PAGI"),
+ "/localdata/class.labels.txt",sep=""),quote="\"", stringsAsFactors=FALSE)[1,])
> #identify dysregulated pathways
> result<-PAGI.Main(dataset,class.labels,nperm = 100,p.val.threshold = -1,FDR.threshold = 0.01,
+ gs.size.threshold.min = 25, gs.size.threshold.max = 500 )

[1] "Running PAGI Analysis..."

> #print the summary results of top ten pathways to screen
> result[[1]][1:10,]

                                Pathway Name SIZE     PathwayID Pathway Score
1                       ErbB signaling pathway   73 path:hsa04012       0.53419
2                   Calcium signaling pathway  146 path:hsa04020       0.49004
3       Phosphatidylinositol signaling system   58 path:hsa04070       0.49456
4                                 Cell cycle   99 path:hsa04110       0.48449
5                       p53 signaling pathway   49 path:hsa04115       0.73209
6                                  Apoptosis   73 path:hsa04210       0.52263
7                               Gap junction   63 path:hsa04540       0.48412
8        Toll-like receptor signaling pathway   83 path:hsa04620       0.49664
9       RIG-I-like receptor signaling pathway   51 path:hsa04622       0.50295
10 Natural killer cell mediated cytotoxicity   92 path:hsa04650       0.48677
   NOM p-val FDR q-val Tag \\% Gene \\% Signal
1          0         0   0.151   0.0658  0.142
2          0         0    0.26    0.145  0.226
3          0         0   0.259    0.104  0.233
4          0         0   0.333    0.273  0.245
5          0         0   0.184   0.0358  0.178
6          0         0   0.164   0.0688  0.154
7          0         0   0.286    0.138  0.248
8          0         0   0.301    0.172  0.251
9          0         0   0.294    0.172  0.245
10         0         0   0.446    0.325  0.304

> #print the detail results of top ten genes in the first pathway to screen
> result[[2]][[1]][1:10,]

    # GENE SYMBOL LIST LOC   Tscore(p-value)   GIF CORE_ENRICHMENT
1   1      CDKN1A        1 5.96 ( 1.44e-07 ) 0.967            YES
2   2         SRC       49  2.53 ( 0.00737 ) 0.512            YES
3   3      CAMK2A       75  2.49 ( 0.00814 ) 0.421            YES
4   4      MAP2K1       82  2.45 ( 0.00899 ) 0.419            YES
```

```
5   5      PIK3CA      118   2.08 ( 0.0214 ) 0.606            YES
6   6      MAP2K7      132   2.16 ( 0.0179 )   0.5            YES
7   7      CAMK2B      189   2.18 ( 0.0171 ) 0.375            YES
8   8       PLCG2      246   1.96 ( 0.0279 ) 0.493            YES
9   9        NRAS      461   1.82 ( 0.0375 ) 0.432            YES
10 10        PAK3      644    1.8 ( 0.0391 )  0.32            YES
```

```
> #write the summary results of pathways to tab delimited file.
> write.table(result[[1]], file = "SUMMARY RESULTS.txt", quote=F, row.names=F, sep = "\t")
> #write the detail results of genes for each pathway with FDR.threshold< 0.01 to tab delimited file.
> for(i in 1:length(result[[2]])){
+ gene.report<-result[[2]][[i]]
+ filename <- paste(names(result[[2]][i]),".txt", sep="", collapse="")
+ write.table(gene.report, file = filename, quote=F, row.names=F, sep = "\t")
+ }
```

3 Session Info

The script runs within the following session:

```
R version 2.15.2 (2012-10-26)
Platform: i386-w64-mingw32/i386 (32-bit)

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=Chinese_People's Republic of China.936
[3] LC_MONETARY=Chinese_People's Republic of China.936
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese_People's Republic of China.936

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] Matrix_1.0-9    lattice_0.20-10 PAGI_1.0        igraph_0.6-3

loaded via a namespace (and not attached):
[1] grid_2.15.2  tools_2.15.2
```

References

[Li *et al.*, 2009] Li, C., et al. (2009) Subpathwayminer: A Software Package for Flexible Identification of Pathways. Nucleic Acids Res, 37, e131.

[Subramanian *et al.*, 2005] Subramanian, A., et al. (2008) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102, 15545-15550.