Bioconductor's SPIA package

Adi L. Tarca[1,2,3], Purvesh Khatri[1] and Sorin Draghici[1]

October 28, 2009

[1]Department of Computer Science, Wayne State University
[2]Bioinformatics and Computational Biology Unit of the NIH Perinatology Research Branch
[3]Center for Molecular Medicine and Genetics, Wayne State University

1 Overview

This package implements the Signaling Pathway Impact Analysis (SPIA) algorithm described in Tarca et al. (2009), Khatri et al. (2007) and Draghici et al. (2007). SPIA uses the information from a set of differentially expressed genes and their fold changes, as well as pathways topology in order to assess the significance of the pathways in the condition under the study. The current version of SPIA algorithm uses KEGG signaling pathway data. SPIA ready KEGG pathway data for homo sapiens is included in the package and also available at
`http://bioinformaticsprb.med.wayne.edu/SPIA/`.
The pathways included for each organism are those containing only directed relations between genes/proteins and no reactions.

2 Pathway analysis with SPIA package

This document provides basic introduction on how to use the `SPIA` package. For extended description of the methods used by this package please consult these references: Tarca et al. (2009); Khatri et al. (2007); Draghici et al. (2007).

We demonstrate the functionality of this package using a colorectal cancer dataset obtained using Affymetrix GeneChip technology and available through GEO (GSE4107). The experiment contains 10 normal samples and 12 colorectal cancer samples and is described by Hong et al. (2007). RMA preprocessing of the raw data was performed using the `affy` package, and a two group moderated t-test was applied using the `limma` package. The data frame obtained as an end result from the function topTable in limma is used as starting point for preparing the input data for SPIA. This data frame called `top` was made available in the `colorectalcancer` dataset included in the SPIA package:

```
> library(SPIA)
> data(colorectalcancer)
```

```
> options(digits = 3)
> head(top)

               ID logFC AveExpr    t  P.Value adj.P.Val    B
10738   201289_at  5.96    6.23 23.9 1.79e-17  9.78e-13 25.4
18604   209189_at  5.14    7.49 17.4 1.56e-14  2.84e-10 21.0
11143 201694_s_at  4.15    7.04 16.5 5.15e-14  7.04e-10 20.1
10490 201041_s_at  2.43    9.59 14.1 1.29e-12  1.41e-08 17.7
10913 201464_x_at  1.53    8.22 11.0 1.69e-10  1.15e-06 13.6
11463   202014_at  1.43    5.33 10.5 4.27e-10  2.42e-06 12.8
```

For SPIA to work, we need a vector with log2 fold changes between the two groups for all the genes considered to be differentially expressed. The names of this vector must be Entrez gene IDs. The following lines will add one additional column in the `top` data frame annotating each affymetrix probeset to an Entrez ID. Since there may be several probesets for the same Entrez ID, there are two easy ways to obtain one log fold change per gene. The first option is to use the fold change of the most significant probeset for each gene, while the second option is to average the log fold-changes of all probestes of the same gene. In the example below we used the former approach. The genes in this example are called differentially expressed provided that their FDR p-value is less than 0.05. The following lines start with the `top` data frame and produce two vectors that are required as input by `spia` function:

```
> library(hgu133plus2.db)
> x <- hgu133plus2ENTREZID
> top$ENTREZ <- unlist(as.list(x[top$ID]))
> top <- top[!is.na(top$ENTREZ), ]
> top <- top[!duplicated(top$ENTREZ), ]
> tg1 <- top[top$adj.P.Val < 0.05, ]
> DE_Colorectal = tg1$logFC
> names(DE_Colorectal) <- as.vector(tg1$ENTREZ)
> ALL_Colorectal = top$ENTREZ
```

The `DE_Colorectal` is a vector containing the log2 fold changes of the genes found to be differentially expressed between cancer and normal samples, and `ALL_Colorectal` is a vector with the Entrez IDs of all genes profiled on the microarray. The names of the `DE_Colorectal` are the Entrez gene IDs corresponding to the computed log fold-changes.

```
> DE_Colorectal[1:10]

 3491  2353  1958  1843  3725 23645  9510 84869  7432  1490
 5.96  5.14  4.15  2.43  1.53  1.43  3.94 -1.15  4.72  3.45

> ALL_Colorectal[1:10]

 [1] "3491"  "2353"  "1958"  "1843"  "3725"  "23645" "9510"  "84869" "7432"
[10] "1490"
```

The SPIA algorithm takes as input the two vectors above and produces a table of pathways ranked from the most to the least significant. This can be achieved by calling the `spia` function as follows:

```
> res = spia(de = DE_Colorectal, all = ALL_Colorectal, organism = "hsa",
+     nB = 2000, plots = FALSE, beta = NULL)

Done pathway 1 : PPAR signaling pathway..
Done pathway 2 : MAPK signaling pathway..
Done pathway 3 : ErbB signaling pathway..
Done pathway 4 : Calcium signaling pathway..
Done pathway 5 : Cytokine-cytokine recepto..
Done pathway 6 : Chemokine signaling pathw..
Done pathway 7 : Neuroactive ligand-recept..
Done pathway 8 : Cell cycle..
Done pathway 9 : p53 signaling pathway..
Done pathway 10 : SNARE interactions in ves..
Done pathway 11 : Regulation of autophagy..
Done pathway 12 : mTOR signaling pathway..
Done pathway 13 : Apoptosis..
Done pathway 14 : Vascular smooth muscle co..
Done pathway 15 : Wnt signaling pathway..
Done pathway 16 : Dorso-ventral axis format..
Done pathway 17 : Notch signaling pathway..
Done pathway 18 : Hedgehog signaling pathwa..
Done pathway 19 : TGF-beta signaling pathwa..
Done pathway 20 : Axon guidance..
Done pathway 21 : VEGF signaling pathway..
Done pathway 22 : Focal adhesion..
Done pathway 23 : ECM-receptor interaction..
Done pathway 24 : Cell adhesion molecules (..
Done pathway 25 : Adherens junction..
Done pathway 26 : Tight junction..
Done pathway 27 : Gap junction..
Done pathway 28 : Complement and coagulatio..
Done pathway 29 : Antigen processing and pr..
Done pathway 30 : Toll-like receptor signal..
Done pathway 31 : RIG-I-like receptor signa..
Done pathway 32 : Jak-STAT signaling pathwa..
Done pathway 33 : Natural killer cell media..
Done pathway 34 : T cell receptor signaling..
Done pathway 35 : B cell receptor signaling..
Done pathway 36 : Fc epsilon RI signaling p..
Done pathway 37 : Fc gamma R-mediated phago..
Done pathway 38 : Leukocyte transendothelia..
Done pathway 39 : Circadian rhythm - mammal..
Done pathway 40 : Long-term potentiation..
```

```
Done pathway 41 : Neurotrophin signaling pa..
Done pathway 42 : Long-term depression..
Done pathway 43 : Olfactory transduction..
Done pathway 44 : Taste transduction..
Done pathway 45 : Regulation of actin cytos..
Done pathway 46 : Insulin signaling pathway..
Done pathway 47 : GnRH signaling pathway..
Done pathway 48 : Progesterone-mediated ooc..
Done pathway 49 : Melanogenesis..
Done pathway 50 : Adipocytokine signaling p..
Done pathway 51 : Type II diabetes mellitus..
Done pathway 52 : Type I diabetes mellitus..
Done pathway 53 : Maturity onset diabetes o..
Done pathway 54 : Alzheimer's disease..
Done pathway 55 : Parkinson's disease..
Done pathway 56 : Amyotrophic lateral scler..
Done pathway 57 : Huntington's disease..
Done pathway 58 : Prion diseases..
Done pathway 59 : Vibrio cholerae infection..
Done pathway 60 : Epithelial cell signaling..
Done pathway 61 : Pathogenic Escherichia co..
Done pathway 62 : Pathways in cancer..
Done pathway 63 : Colorectal cancer..
Done pathway 64 : Renal cell carcinoma..
Done pathway 65 : Pancreatic cancer..
Done pathway 66 : Endometrial cancer..
Done pathway 67 : Glioma..
Done pathway 68 : Prostate cancer..
Done pathway 69 : Thyroid cancer..
Done pathway 70 : Basal cell carcinoma..
Done pathway 71 : Melanoma..
Done pathway 72 : Bladder cancer..
Done pathway 73 : Chronic myeloid leukemia..
Done pathway 74 : Acute myeloid leukemia..
Done pathway 75 : Small cell lung cancer..
Done pathway 76 : Non-small cell lung cance..
Done pathway 77 : Asthma..
Done pathway 78 : Autoimmune thyroid diseas..
Done pathway 79 : Systemic lupus erythemato..
Done pathway 80 : Allograft rejection..
Done pathway 81 : Graft-versus-host disease..
Done pathway 82 : Arrhythmogenic right vent..
Done pathway 83 : Dilated cardiomyopathy..
```

```
> res$Name = substr(res$Name, 1, 10)
> res[1:15, -12]
```

```
          Name    ID pSize NDE     tA     pNDE    pPERT       pG    pGFdr
1  Parkinson' 05012   104  56 -12.04 2.15e-14 0.042000 3.21e-14 2.54e-12
2  Alzheimer' 05010   146  69  -6.22 1.15e-13 0.232000 8.58e-13 3.39e-11
3  Focal adhe 04510   172  63  97.79 3.08e-07 0.000005 4.35e-11 1.14e-09
4  Huntington 05016   163  65  -3.19 4.43e-09 0.187000 1.81e-08 3.58e-07
5  ECM-recept 04512    74  26  21.95 1.76e-03 0.000005 1.72e-07 2.71e-06
6  PPAR signa 03320    64  30  -3.14 1.21e-06 0.066000 1.39e-06 1.83e-05
7  Axon guida 04360   119  47   9.27 8.12e-07 0.338000 4.42e-06 4.99e-05
8  Circadian  04710     8   6  -3.15 1.25e-03 0.137000 1.65e-03 1.63e-02
9  Colorectal 05210    77  26   7.09 3.31e-03 0.059000 1.86e-03 1.63e-02
10 Small cell 05222    76  21  24.74 6.93e-02 0.005000 3.11e-03 2.37e-02
11 Wnt signal 04310   140  43  -8.08 1.76e-03 0.226000 3.50e-03 2.37e-02
12 Regulation 04810   187  55  15.29 1.38e-03 0.297000 3.60e-03 2.37e-02
13 MAPK signa 04010   248  71   3.40 7.14e-04 0.666000 4.11e-03 2.50e-02
14 Renal cell 05211    62  21  -8.05 7.51e-03 0.090000 5.61e-03 3.17e-02
15 Pathogenic 05130    45  13  17.34 1.01e-01 0.018000 1.33e-02 6.98e-02
     pGFWER    Status
1  2.54e-12 Inhibited
2  6.77e-11 Inhibited
3  3.43e-09 Activated
4  1.43e-06 Inhibited
5  1.36e-05 Activated
6  1.10e-04 Inhibited
7  3.49e-04 Activated
8  1.30e-01 Inhibited
9  1.47e-01 Activated
10 2.46e-01 Activated
11 2.77e-01 Inhibited
12 2.84e-01 Activated
13 3.25e-01 Activated
14 4.43e-01 Inhibited
15 1.00e+00 Activated
```

If the `plots` argument is set to `TRUE` in the function call above, a plot like the one shown in Figure 1 is produced for each pathway on which there are differentially expressed genes. These plots are saved in a pdf file in the current directory.

An overall picture of the pathways significance according to both the over-representation evidence and perturbations based evidence can be obtained with the function `plotP` and shown in Figure 2. In this plot, the horizontal axis represents the p-value (minus log of) corresponding to the probability of obtaining at least the observed number of genes (NDE) on the given pathway just by chance. The vertical axis represents the p-value (minus log of) corresponding to the probability of obtaining the observed total accumulation (tA) or more extreme on the given pathway just by chance. The computation of pPERT is described in Tarca et al. (2009). In Figure 2 each pathway is shown as a bullet point, and those significant at 5% (set by the `threshold` argument in `plotP`) after Bonferroni correction are shown in red.

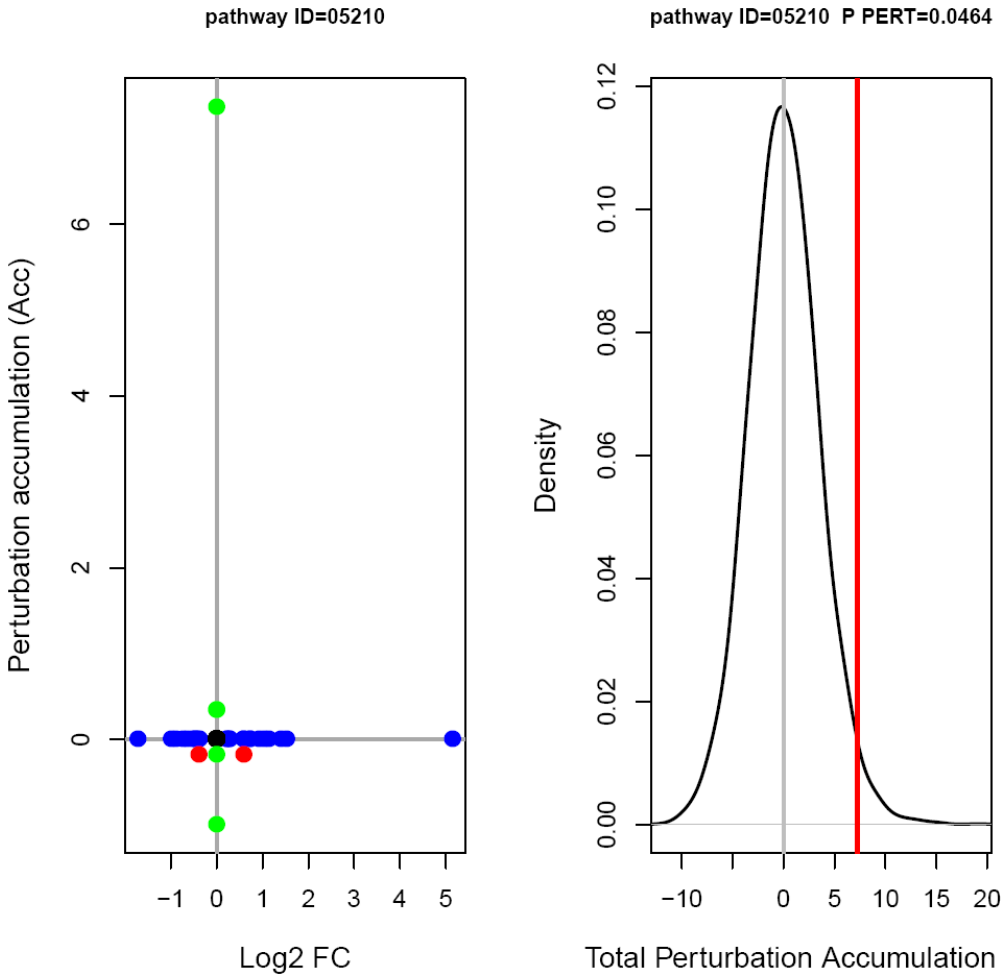

Figure 1: Perturbations plot for colorectal cancer pathway (KEGG ID hsa:05210) using the `colorectalcancer` dataset. The perturbation of all genes in the pathway are shown as a function of their initial log2 fold changes (left panel). Non DE genes are assigned 0 log2 fold-change. The null distribution of the net accumulated perturbations is also given (right panel). The observed net accumulation tA with the real data is shown as a red vertical line.

```
> plotP(res, threshold = 0.05)
```

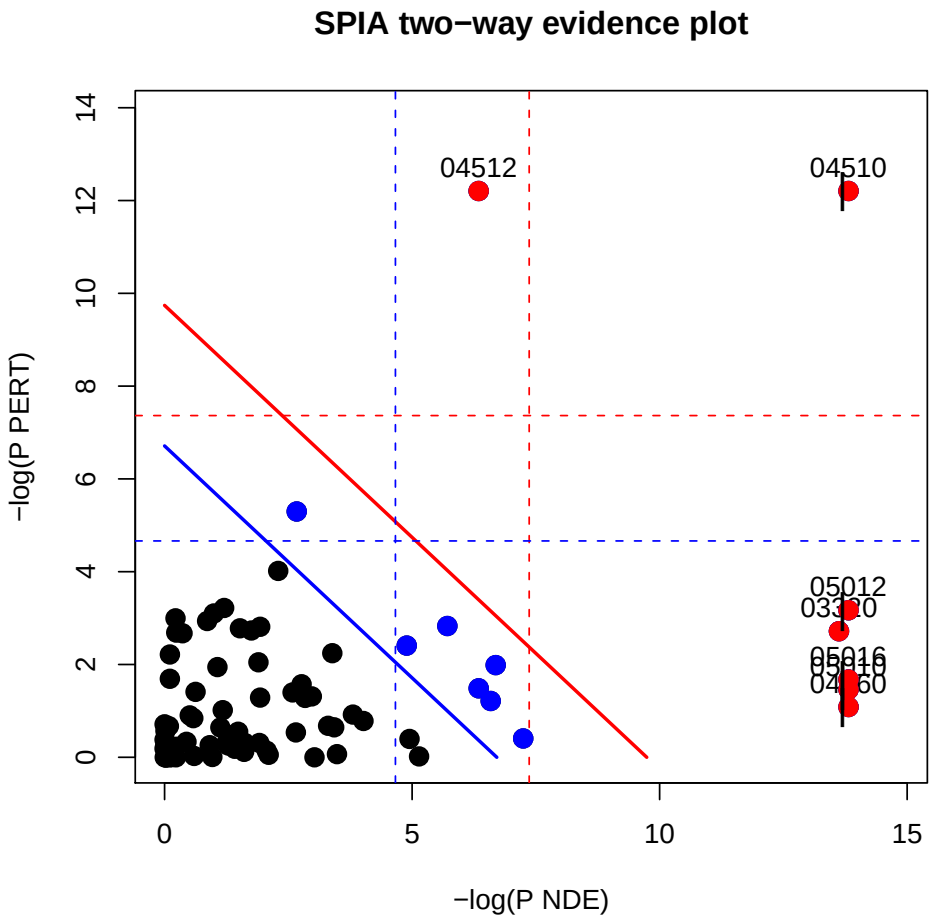

Figure 2: SPIA evidence plot for the colorectal cancer dataset. Each pathway is represented by one dot. The pathways at the right of the red oblique line are significant after Bonferroni correction of the global p-values, pG. The pathways at the right of the blue oblique line are significant after a FDR correction of the global p-values, pG.

SPIA algorithm is illustrated also using the Vessels dataset:

```
> data(Vessels)
> res <- spia(de = DE_Vessels, all = ALL_Vessels, organism = "hsa",
+     nB = 500, plots = FALSE, beta = NULL, verbose = FALSE)
> res$Name = substr(res$Name, 1, 10)
> res[1:15, -12]
```

```
         Name    ID pSize NDE       tA     pNDE pPERT       pG  pGFdr pGFWER
1  Axon guida 04360   128  12 -5.88523 0.000208 0.104 0.000254 0.0188 0.0188
2  Focal adhe 04510   191  15 -5.03618 0.000257 0.412 0.001075 0.0267 0.0795
3  Neuroactiv 04080   255  18 -0.51036 0.000247 0.432 0.001081 0.0267 0.0800
4  Regulation 04810   203  13  9.55606 0.004019 0.052 0.001980 0.0366 0.1465
5  Notch sign 04330    46   4  7.39960 0.036603 0.008 0.002675 0.0396 0.1980
6  Complement 04610    67   7  4.85746 0.002325 0.348 0.006570 0.0680 0.4862
7  Graft-vers 05332    41   6  0.00000 0.000813 1.000 0.006597 0.0680 0.4882
8      Asthma 05310    29   5  0.00000 0.001038 1.000 0.008167 0.0680 0.6043
9  Type I dia 04940    43   6  0.00000 0.001053 1.000 0.008270 0.0680 0.6120
10 Antigen pr 04612    88   7  1.54604 0.010418 0.124 0.009884 0.0731 0.7314
11 Wnt signal 04310   151  11  1.16119 0.002986 0.624 0.013573 0.0913 1.0000
12 Allograft  05330    37   5  0.00000 0.003183 1.000 0.021483 0.1325 1.0000
13 Epithelial 05120    67   5  2.00907 0.036325 0.192 0.041606 0.2367 1.0000
14 ECM-recept 04512    83   7 -0.00171 0.007649 0.996 0.044777 0.2367 1.0000
15 Leukocyte  04670   109   8 -0.59837 0.010167 0.868 0.050567 0.2495 1.0000
      Status
1  Inhibited
2  Inhibited
3  Inhibited
4  Activated
5  Activated
6  Activated
7  Inhibited
8  Inhibited
9  Inhibited
10 Activated
11 Activated
12 Inhibited
13 Activated
14 Inhibited
15 Inhibited
```

The pathway image as provided by KEGG having the differentially expressed genes highlighted in red can be obtained by pasting in a web browser the links available in the KEGGLINK column of the data frame produced by the function spia. For example,

```
> res[, "KEGGLINK"][20]
```

```
[1] "http://www.genome.jp/dbget-bin/show_pathway?hsa05211+7424+6513+112399+5155+3725"
```

is the link that would display the image of the 20th pathway in the res dataframe above.
Note that the results for these datasets my differ from the ones described in Tarca et al. (2009) since a) the pathways database used herein was updated and b) the default beta values were changed.
The directed adjacency matrices of the graphs describing the different types of relations between genes/proteins (such as activation or repression) used by SPIA are available in the `extdata/hsaSPIA.RData` file for the homo sapiens organism. The types of relations considered by SPIA and the default weight (beta coefficient) given to them are:

```
> rel <- c("activation", "compound", "binding/association", "expression",
+     "inhibition", "activation_phosphorylation", "phosphorylation",
+     "indirect", "inhibition_phosphorylation", "dephosphorylation_inhibition",
+     "dissociation", "dephosphorylation", "activation_dephosphorylation",
+     "state", "activation_indirect", "inhibition_ubiquination",
+     "ubiquination", "expression_indirect", "indirect_inhibition",
+     "repression", "binding/association_phosphorylation", "dissociation_phosphorylation",
+     "indirect_phosphorylation")
> beta = c(1, 0, 0, 1, -1, 1, 0, 0, -1, -1, 0, 0, 1, 0, 1, -1,
+     0, 1, -1, -1, 0, 0, 0)
> names(beta) <- rel
> cbind(beta)
```

	beta
activation	1
compound	0
binding/association	0
expression	1
inhibition	-1
activation_phosphorylation	1
phosphorylation	0
indirect	0
inhibition_phosphorylation	-1
dephosphorylation_inhibition	-1
dissociation	0
dephosphorylation	0
activation_dephosphorylation	1
state	0
activation_indirect	1
inhibition_ubiquination	-1
ubiquination	0
expression_indirect	1
indirect_inhibition	-1
repression	-1
binding/association_phosphorylation	0
dissociation_phosphorylation	0
indirect_phosphorylation	0

A 0 value for a given relation type results in discarding those type of relations from the analysis for all pathways. The default values of `beta` can changed by the user at any time by setting the `beta` argument of the `spia` function call.
Other organisms' KEGG pathway data can be downloaded from `http://bioinformaticsprb.med.wayne.edu/SPIA` as a "[org]SPIA.RData" file and copied into the `extdata` directory of the SPIA package, and therefore make it available to the function `spia`.
The user has the ability to generate his own gene/protein relation data and put it in a list format as the one shown in the hsaSPIA.RData file. In this file, each pathway data is included in a list:

```
> load(file = paste(system.file("extdata/hsaSPIA.RData", package = "SPIA")))
> names(path.info[["05210"]])
```

```
 [1] "activation"                          "compound"
 [3] "binding/association"                 "expression"
 [5] "inhibition"                          "activation_phosphorylation"
 [7] "phosphorylation"                     "indirect"
 [9] "inhibition_phosphorylation"          "dephosphorylation_inhibition"
[11] "dissociation"                        "dephosphorylation"
[13] "activation_dephosphorylation"        "state"
[15] "activation_indirect"                 "inhibition_ubiquination"
[17] "ubiquination"                        "expression_indirect"
[19] "indirect_inhibition"                 "repression"
[21] "binding/association_phosphorylation" "dissociation_phosphorylation"
[23] "indirect_phosphorylation"            "nodes"
[25] "title"                               "NumberOfReactions"
```

```
> path.info[["05210"]][["activation"]][48:60, 55:60]
```

```
     8313 5900 5879 5880 5881 332
369     0    0    0    0    0   0
5894    0    0    0    0    0   0
673     0    0    0    0    0   0
5599    0    0    1    1    1   0
5601    0    0    1    1    1   0
5602    0    0    1    1    1   0
8312    0    0    0    0    0   0
8313    0    0    0    0    0   0
5900    0    0    0    0    0   0
5879    0    1    0    0    0   0
5880    0    1    0    0    0   0
5881    0    1    0    0    0   0
332     0    0    0    0    0   0
```

In the matrix above, only 0 and 1 values are allowed. 1 means the gene/protein given by the column has a relation of type "activation" with the gene/protein given by the row of the matrix.
Using other R packages such as `graph` and `Rgraphviz` one can visualize the richness of gene/protein relations of each type in each pathway. Firstly we load the required packages and create a function that can be used to plot as a graph each type of relation of any pathway, as used by SPIA.

```
> library(graph)
> library(Rgraphviz)
> plotG <- function(B) {
+     nnms <- NULL
+     colls <- NULL
+     mynodes <- colnames(B)
+     L <- list()
+     n <- dim(B)[1]
+     for (i in 1:n) {
+         L[i] <- list(edges = rownames(B)[abs(B[, i]) > 0])
+         if (sum(B[, i] != 0) > 0) {
+             nnms <- c(nnms, paste(colnames(B)[i], rownames(B)[B[,
+                 i] != 0], sep = "~"))
+         }
+     }
+     names(L) <- rownames(B)
+     g <- new("graphNEL", nodes = mynodes, edgeL = L, edgemode = "directed")
+     plot(g)
+ }
```

We plot then the "activation" relations in the ErbB signaling pathway, based on the `hsaSPIA` data. For more details on how to use the main function in this package use "?spia".

References

S. Draghici, P. Khatri, A. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17, 2007.

Y. Hong, K. S. Ho, K. W. Eu, and P. Y. Cheah. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res*, 13(4):1107–14, 2007.

P. Khatri, S. Draghici, A. L. Tarca, S. S. Hassan, and R. Romero. A system biology approach for the steady-state analysis of gene signaling networks. In *12th Iberoamerican Congress on Pattern Recognition*, Valparaiso, Chile, November 13-16 2007.

A. L. Tarca, S. Draghici, P. Khatri, S. Hassan, P. Mital, J. Kim, C. Kim, J. P. Kusanovic, and R. Romero. A signaling pathway impact analysis for microarray experiments. *Bioinformatics*, 25:75–82, 2009.

```
> plotG(path.info[["04012"]][["activation"]])
```

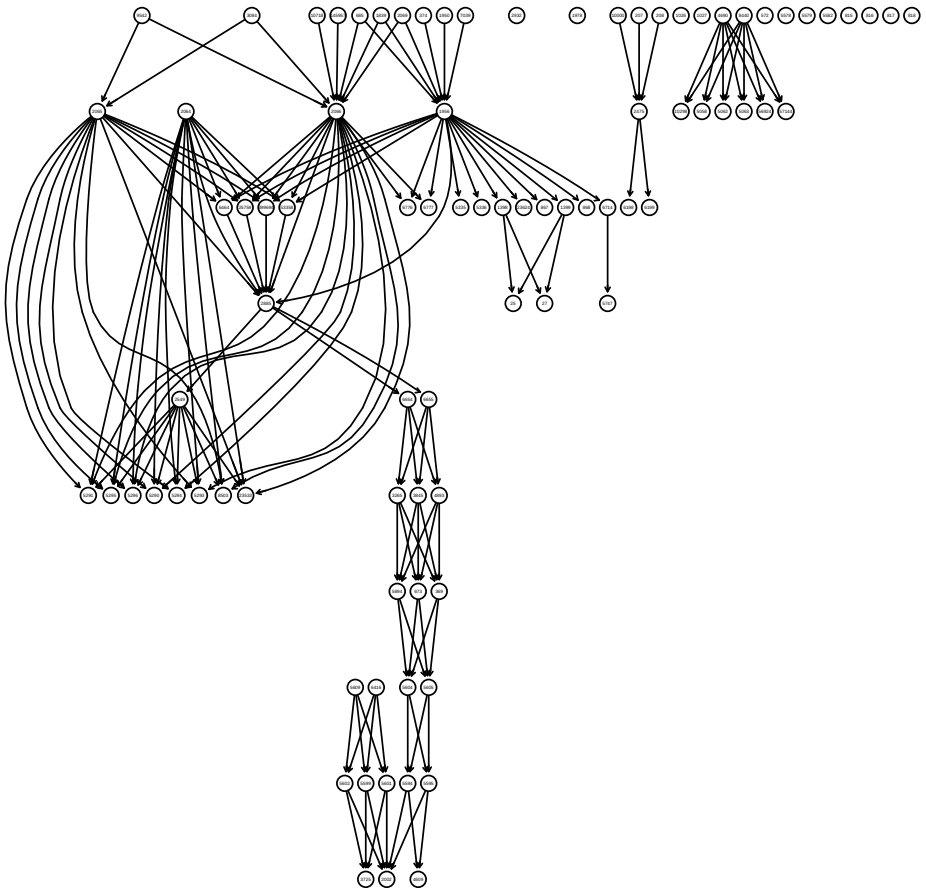

Figure 3: Display of the "activation" relations in the ErbB signaling pathway, based on the hsaSPIA data.