Bioconductor's SPIA package

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

October 15, 2010

[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.
The KEGG data that was preprocessed for SPIA analysis was downloaded from KEGG's ft repository on: 09/22/2010.

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"  "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 = FA
```

```
Done pathway 1 : PPAR signaling pathway..
Done pathway 2 : MAPK signaling pathway..
Done pathway 3 : ErbB signaling pathway..
...
Done pathway 100 : Dilated cardiomyopathy..
Done pathway 101 : Viral myocarditis..
```

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

	Name	ID	pSize	NDE	tA	pNDE	pPERT	pG	pGFdr	pGFWER	Status
1	Parkinson'	05012	106	56	-12.026	7.22e-14	0.063000	1.55e-13	1.50e-11	1.50e-11	Inhibited
2	Alzheimer'	05010	146	69	-6.240	1.33e-13	0.241000	1.03e-12	4.99e-11	9.99e-11	Inhibited
3	Focal adhe	04510	177	63	100.415	1.09e-06	0.000005	1.47e-10	4.74e-09	1.42e-08	Activated
4	Huntington	05016	164	65	-3.273	6.68e-09	0.179000	2.58e-08	6.25e-07	2.50e-06	Inhibited
5	ECM-recept	04512	74	26	22.173	1.84e-03	0.000005	1.79e-07	3.47e-06	1.74e-05	Activated
6	PPAR signa	03320	64	30	-3.099	1.30e-06	0.051000	1.16e-06	1.87e-05	1.12e-04	Inhibited
7	Axon guida	04360	119	47	9.220	8.87e-07	0.341000	4.84e-06	6.71e-05	4.70e-04	Activated
8	Small cell	05222	75	21	25.070	6.30e-02	0.003000	1.81e-03	2.19e-02	1.75e-01	Activated
9	Wnt signal	04310	138	43	-8.449	1.36e-03	0.188000	2.38e-03	2.56e-02	2.31e-01	Inhibited
10	Regulation	04810	192	56	15.464	1.66e-03	0.273000	3.95e-03	3.83e-02	3.83e-01	Activated
11	Lysosome	04142	116	36	-0.753	3.45e-03	0.161000	4.72e-03	4.16e-02	4.58e-01	Inhibited
12	MAPK signa	04010	245	69	5.732	1.47e-03	0.419000	5.16e-03	4.17e-02	5.00e-01	Activated
13	Renal cell	05211	64	21	-7.963	1.15e-02	0.095000	8.56e-03	6.38e-02	8.30e-01	Inhibited
14	Pathogenic	05130	48	13	17.180	1.52e-01	0.018000	1.89e-02	1.31e-01	1.00e+00	Activated
15	Circadian	04710	16	8	-2.640	7.23e-03	0.410000	2.02e-02	1.31e-01	1.00e+00	Inhibited

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. The Colorectal cancer pathway is shown in green.
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.

SPIA algorithm is illustrated also using the Vessels dataset:

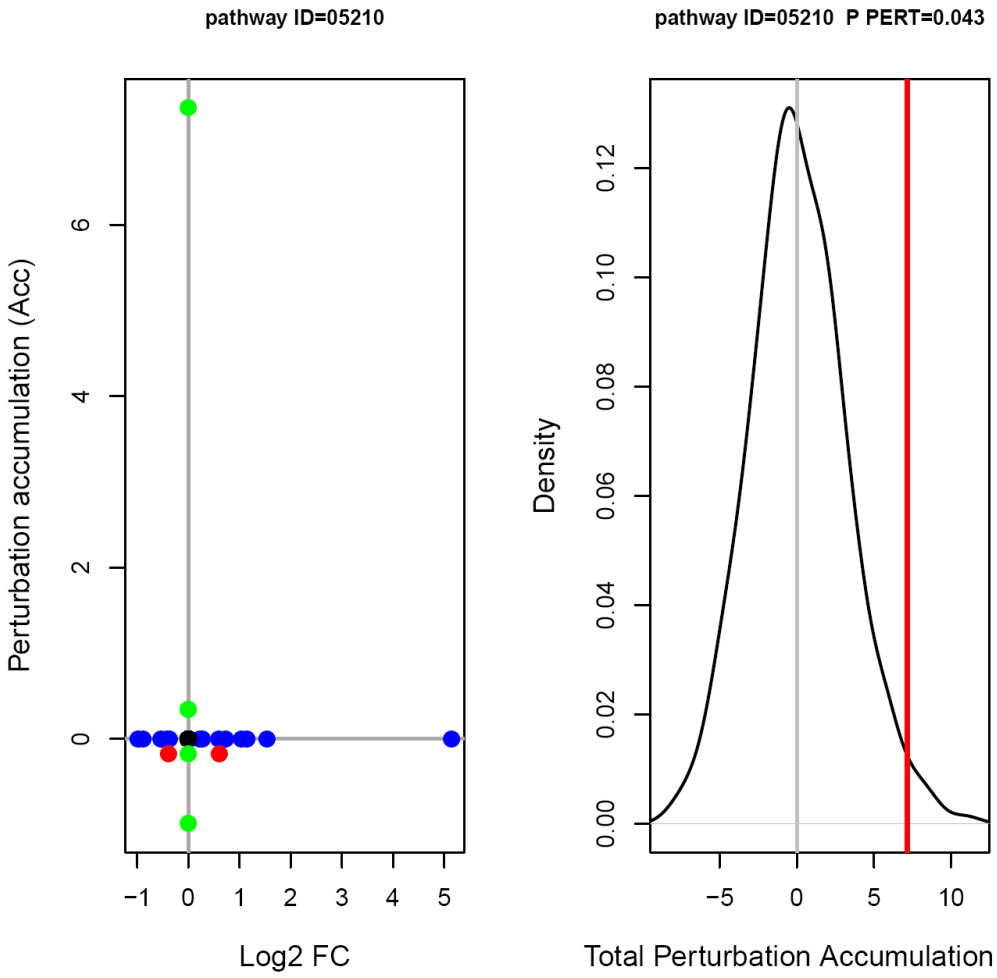

Figure 1: Perturbations plot for colorectal cancer pathway (KEGG ID hsa:05210) using the `colorectalcancer` dataset. The perturbation of all genes in the pathway are shown as a function of their initial log2 fold changes (left panel). Non DE genes are assigned 0 log2 fold-change. The null distribution 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)
> points(I(-log(pPERT)) ~ I(-log(pNDE)), data = res[res$ID == "05210", ], col = "green", pch =
```

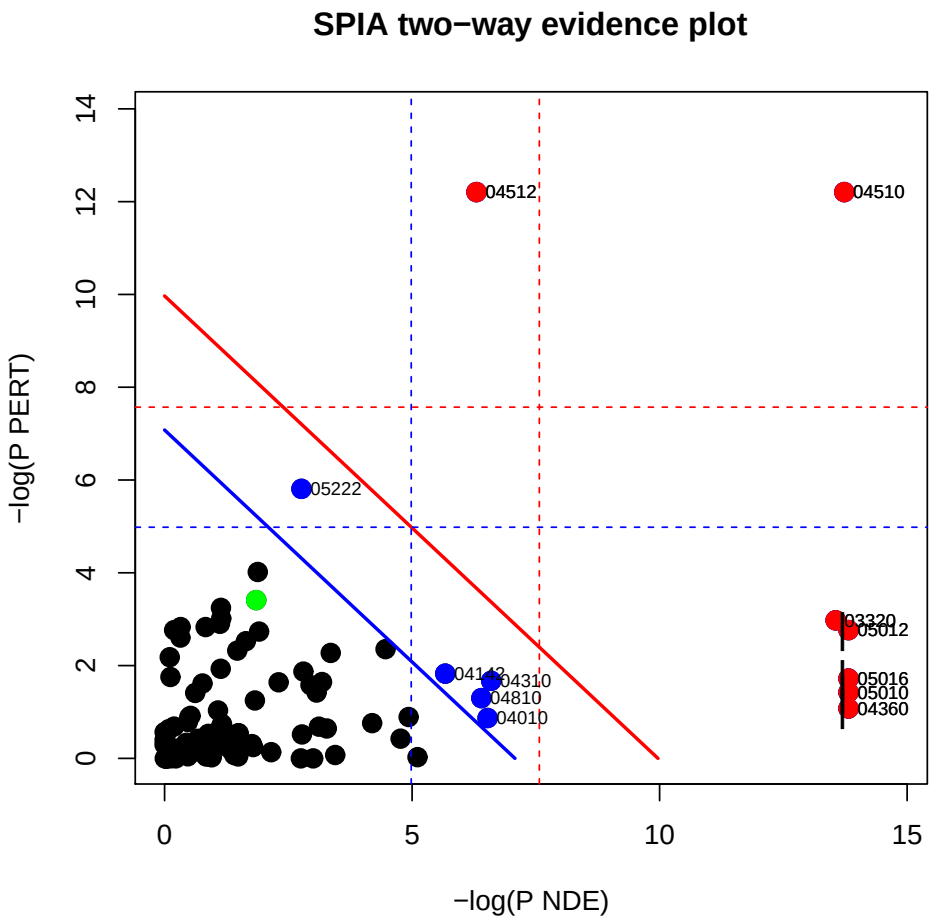

Figure 2: SPIA evidence plot for the colorectal cancer dataset. Each pathway is represented by one dot. The pathways at the right of the red oblique line are significant after Bonferroni correction of the global p-values, pG. The pathways at the right of the blue oblique line are significant after a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

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

	Name	ID	pSize	NDE	tA	pNDE	pPERT	pG	pGFdr	pGFWER	Status
1	Axon guida	04360	128	12	-5.7374	0.000208	0.088	0.000218	0.0194	0.0194	Inhibited
2	Focal adhe	04510	198	16	-6.2100	0.000116	0.372	0.000478	0.0213	0.0425	Inhibited
3	Regulation	04810	210	14	9.4361	0.002001	0.064	0.001276	0.0325	0.1135	Activated
4	Viral myoc	05416	70	8	-1.6656	0.000635	0.288	0.001757	0.0325	0.1564	Inhibited
5	Neuroactiv	04080	271	18	-0.5104	0.000512	0.428	0.002066	0.0325	0.1839	Inhibited
6	Intestinal	04672	46	7	0.0000	0.000234	1.000	0.002193	0.0325	0.1952	Inhibited
7	Antigen pr	04612	76	7	1.4922	0.004739	0.088	0.003662	0.0466	0.3260	Activated
8	Leishmania	05140	68	8	0.0122	0.000522	0.992	0.004432	0.0493	0.3944	Activated
9	Complement	04610	67	7	4.5721	0.002325	0.344	0.006504	0.0587	0.5788	Activated
10	Graft-vers	05332	41	6	0.0000	0.000813	1.000	0.006597	0.0587	0.5871	Inhibited
11	Asthma	05310	29	5	0.0000	0.001038	1.000	0.008167	0.0613	0.7268	Inhibited
12	Type I dia	04940	43	6	0.0000	0.001053	1.000	0.008270	0.0613	0.7360	Inhibited
13	Notch sign	04330	46	4	7.5682	0.036603	0.032	0.009077	0.0621	0.8079	Activated
14	Wnt signal	04310	150	11	1.0119	0.002836	0.684	0.014055	0.0893	1.0000	Activated
15	Leukocyte	04670	116	9	-1.9287	0.004608	0.616	0.019484	0.1156	1.0000	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?hsa04540+3356+983+6714+5155+80310"
```

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", "activa
+     "phosphorylation", "indirect", "inhibition_phosphorylation", "dephosphorylation_inhibitio
+     "activation_dephosphorylation", "state", "activation_indirect", "inhibition_ubiquination"
+     "indirect_inhibition", "repression", "binding/association_phosphorylation", "dissociation
> 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/SPI` 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"                        "binding/assoc
 [4] "expression"                      "inhibition"                      "activation_ph
 [7] "phosphorylation"                 "indirect"                        "inhibition_ph
[10] "dephosphorylation_inhibition"    "dissociation"                    "dephosphoryla
[13] "activation_dephosphorylation"    "state"                           "activation_in
[16] "inhibition_ubiquination"         "ubiquination"                    "expression_in
[19] "indirect_inhibition"             "repression"                      "binding/assoc
[22] "dissociation_phosphorylation"    "indirect_phosphorylation"        "nodes"
[25] "title"                           "NumberOfReactions"
```

```
> path.info[["05210"]][["activation"]][25:35, 30:40]
```

	5602	8312	8313	5900	387	5879	5880	5881	332	4609	595
369	0	0	0	0	0	0	0	0	0	0	0
5894	0	0	0	0	0	0	0	0	0	0	0
673	0	0	0	0	0	0	0	0	0	0	0
5599	0	0	0	0	1	1	1	1	0	0	0
5601	0	0	0	0	1	1	1	1	0	0	0
5602	0	0	0	0	1	1	1	1	0	0	0
8312	0	0	0	0	0	0	0	0	0	0	0
8313	0	0	0	0	0	0	0	0	0	0	0
5900	0	0	0	0	0	0	0	0	0	0	0
387	0	0	0	1	0	0	0	0	0	0	0
5879	0	0	0	1	0	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.

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

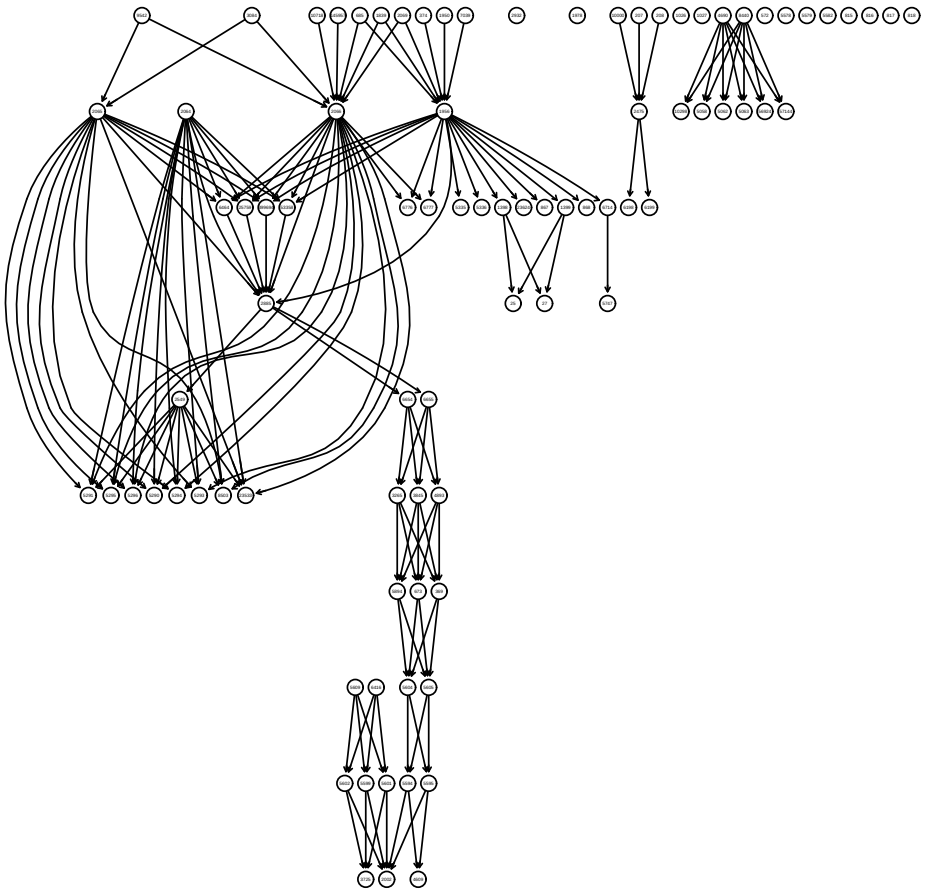

Figure 3: Display of the "activation" relations in the ErbB signaling pathway, based on the hsaSPIA data.

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.