IMPROVING THE OUTPUT OF SIGNALING PATHWAY IMPACT ANALYSIS

Associate Professor PhD. Mohammad Ohid ULLAH
Shahjalal University of Science and Technology, Bangladesh

Abstract

Recently biologists want to find out the significant gene sets instead of individual gene. Many packages in different software mostly in R language were developed to find out the gene sets which are significantly regulated. Among them some packages are able to discover up and down regulation as well. Signaling pathway impact analysis (SPIA) is one of them. In this study an approach is mentioned which can be improved the output of SPIA. I proposed that using moderated t values gives more significant results instead of using logFC (log fold change) from Limma's output in order to calculate probability of perturbation in SPIA.

Key words: Microarray, Gene, Signaling pathway, False discovery rate (FDR), Fold changes, and Moderated t value.

***

The microarray experiment is able to compare two groups of samples like diseased (treatment) and normal (control) samples to find out differentially expressed genes. In order to detect significant gene sets or pathways several methods and packages were already developed. The Gene Set Enrichment Analysis (GSEA) proposed by [1] considering the entire distribution of genes rather than individual genes. The GSEA approach is able to identify a significant gene set between the treatment and control samples for which no single gene was found to be differentially expressed using the over representation analysis (ORA) approach. In addition several methods for gene set analysis were developed such as GSA methods [2] for alternatives to the ORA approach, MaxMean statistic [3] for summarizing gene-sets, and a restandardization for more precise inferences, after reducing the dimension of the gene expression data matrix to its first principal component and then applied t-statistic[4], Hotelling's $T2$-statistic [5] with a similar dimensional reduction approach. global test [6] by modeling gene expressions as random effects in a logistic regression model, analysis of covariance (ANCOVA) test [7, 8]. GSA analysis of GO terms [9]. test statistics based on the two-sample t-statistics [10, 11], SAM-GS test [12] based on the SAM statistic, SAM-GS statistic [13] using the structure of regression model and Multivariate analysis of variance (MANOVA) [14] for gene set analysis. The SAM-GS test is the

only GSA method to concentrate on the variability present in microarray gene expression studies by incorporating the SAM constant into the test statistic.

Recently, Tarca *et al.* (2009) developed a novel signaling pathway impact analysis (SPIA) [15] combining probability of differential expressed genes and probability of perturbation. These probabilities can be calculated using a bootstrap approach. This paper describes a method which is based on SPIA. I proposed that the probability of perturbation can be calculated using the moderated t values from limma [16] instead of logFC (log fold change). The moderated t values are considered the variation of the data within treatment and control samples. On the other hand logFC is not able to consider the variation of the data. The small variability can lead to inflation of the *t*-test statistic due to very small denominator, and therefore genes whose average expressions corresponding to the two groups are extremely close can be identified as significant.

METHODS

In SPIA [15], the over-representation of differential expressed genes in a given pathway and the abnormal perturbation of that pathway which measuring the expression changes across the pathway topology. These arguments are captured by independent probability values, P_{NDE} (probability of differential expressed genes) and P_{PERT} (probability of perturbation). Tarca *et al*. (2009) used logFC to find out the probability of perturbation. The proposed method is based on SPIA, just replace moderated t values [14] instead of logFC in order to find out the probability of perturbation and the probability of number of differential expressed gene remain same like SPIA. The logFC are not able to considered variation of the data within each treatment and control experiment but moderated t values are able to considered variation of the data. So the P_{PERT} can be defined as in the following way:
At first, define a gene perturbation factor as:

$$PF(g_i)= t_{gi} + \sum_{j=1}^{n} \beta_{ij} . \frac{PF(g_j)}{N_{ds}(g_j)} \quad (1)$$

In Eq. (1), the term t_{gi} moderated t values which was observed in limma output. The second term in Eq.(1) represents exactly the same as SPIA. Similarly calculate the net perturbation accumulation at the level of each gene, ACCg, the difference between the perturbation factor and the observed moderated t values:

$$Acc(gi)= PF(gi)- t_{gi} \quad (2)$$

The vector of perturbation accumulation ACC can be obtained using the matrix form:

$$Acc=B.(I-B)-1.T \quad (3)$$

Except the T vector the Eq (3) is same as SPIA. The T vector can be obtained the following way:

$$T=\begin{bmatrix} t_{g1} \\ t_{g2} \\ \dots \\ t_{gn} \end{bmatrix} \quad (4)$$

Now following SPIA procedure the total net accumulated perturbation in the pathway is computed as $t_A = \sum_{i=1}^{n} Acc(g_i)$. Therefore the probability of perturbation becomes:

$$P_{PERT}=P(T_A \geq t_A \mid H_0).$$

This probability can also be calculated using a bootstrap approach. The two types of evidence, P_{NDE} and P_{PERT} are finally combined like SPIA in one global probability value, P_G, which is used to rank the pathways and test the research hypothesis that the pathway is significantly perturbed under the condition of study. The $P_G = c_i - c_i.\ln(c_i)$; where $c_i = P_{NDE}(i).P_{PERT}(i)$.

RESULTS AND DISCUSSION

Using the bioconductor's SPIA packages [15] and their colorectal cancer data in this package. The current version of SPIA uses KEGG signaling pathway data. In order to find out the significant pathways in colorectal cancer data at first I used SPIA algorithm where the object *DE_Colerectal=tg1$**logFC*** (http://bioconductor.org /packages/2.4/bioc/vignettes/SPIA/inst/doc/SPIA.pdf). Then I used my proposed approach where I used SPIA package just modifying the object *DE_Colerectal=tg1$**t.*** That is the names of the DE_Colerectal are the Entrez gene IDs corresponding to the computed moderated t-values. In both cases the top 15 pathways results were presented in the following two tables. It was observed that overall the pPERT, pG and pGFdr were decreasing when I used moderated t values compare to logFC (see Table 1 and Table 2).

SPIA results on Colerectal cancer dataset using logFC in P_{PERT}

Table 1

KEGG Pathway	tA	pNDE	pPERT	pG	pGFdr	pGFWER	Status
Parkinson's.. 5012	-12.6049	0.0000	0.03800	0.0000	**0.0000**	0.0000	Inhibited
Alzheimer's.. 5010	-7.06218	0.0000	0.15100	0.0000	**0.0000**	0.0000	Inhibited
Focal adh..4510	62.9004	0.0000	0.0000	0.0000	**0.0000**	0.0000	Activated
ECM-recep..4512	19.7775	0.0024	0.0000	0.0000	**0.0000**	0.0000	Activated
Axon guid..4360	7.6297	0.0000	0.3990	0.0000	**0.0000**	0.0003	Activated
Colorectal..5210	7.2566	0.0035	0.0470	0.0016	**0.0179**	0.1097	Activated
MAPK sig..4010	5.4930	0.0004	0.4850	0.0019	**0.0179**	0.1259	Activated
Wnt sig..4310	-8.2171	0.0019	0.2130	0.0037	**0.0312**	0.2495	Inhibited
Regulation..4810	8.0732	0.0013	0.3610	0.0043	**0.0323**	0.2902	Activated
Renal cell..5211	-7.6921	0.0099	0.0880	0.0071	**0.0484**	0.4836	Inhibited
Dentator..5050	-0.8941	0.0023	0.6290	0.0108	0.0674	0.7421	Inhibited
Notch sig..4330	3.6119	0.0041	0.5100	0.0149	0.0852	1.0000	Activated
Circadian..4710	0.0000	0.0029	1.0000	0.0201	0.0927	1.0000	Inhibited
Tight jun..4530	1.8705	0.0043	0.6820	0.0202	0.0927	1.0000	Activated
Apoptosis..4210	-15.4708	0.0393	0.0750	0.0202	0.0927	1.0000	Inhibited

SPIA results on Colerectal cancer dataset using moderated t values in P_{PERT}

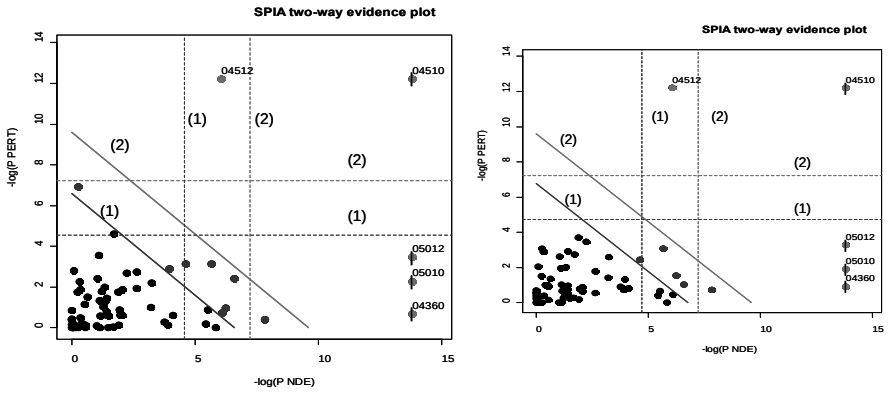
Table 2

KEGG Pathway	tA	pNDE	pPERT	pG	pGFdr	pGFWER	Status
Parkinson's ..5012	-54.3523	0.0000	0.0310	0.0000	**0.0000**	0.0000	Inhibited
Alzheimer's..5010	-33.2952	0.0000	0.1030	0.0000	**0.0000**	0.0000	Inhibited
Focal adh..4510	263.3591	0.0000	0.0000	0.0000	**0.0000**	0.0000	Activated
ECM-recep..4512	73.1511	0.0023	0.0000	0.0000	**0.0000**	0.0001	Activated
Axon guid..4360	24.8177	0.0000	0.5100	0.0000	**0.0001**	0.0003	Activated
Regulation..4810	61.7339	0.0014	0.0890	0.0012	**0.0139**	0.0832	Activated
Colorectal..5210	28.8157	0.0035	0.0440	0.0015	**0.0148**	0.1033	Activated
MAPK sign..4010	14.6213	0.0004	0.6660	0.0024	**0.0209**	0.1671	Activated
Renal cell..5211	-39.6427	0.0099	0.0430	0.0037	**0.0286**	0.2573	Inhibited
Wnt sig..4310	-23.0880	0.0019	0.3850	0.0061	**0.0395**	0.4205	Inhibited
Cytokine-cyto..4060	51.3352	0.7700	0.0010	0.0063	**0.0395**	0.4340	Activated
Gap junction..4540	78.4665	0.0189	0.0550	0.0082	**0.0459**	0.5658	Activated
Dentator..5050	-5.6547	0.0023	0.4890	0.0086	**0.0459**	0.5961	Inhibited
Notch sig..4330	18.5283	0.0041	0.4130	0.0123	0.0602	0.8516	Activated
Melanoma..5218	123.3965	0.1783	0.0100	0.0131	0.0602	0.9016	Activated

A significance threshold of 5% was used on the False Discovery Rate (FDR) corrected p-values in order to detect pathway significance. Using logFC, it was found that 10 pathways are significant but using moderated t values the number of significant pathways are 13. That is applying my proposed approach improving the outcome of SPIA . This is because of variation of the

data. The t values capture the variation of the data within experiment while logFC can't able to capture the variation of the data. The path ways *Cytokine-cytokine receptor*, *Gap junction* and int *Dentatorubropallidoluysian atr* show significant considering the t values at the object DE_colerectal in SPIA.

The SPIA two way evidence plots also indicate that using moderated t values in P_{PERT} gives more number of significant pathways.



(a) (b)

Fig. 1. SPIA two-way evidence plot for the colectal cancer dataset. Panel (a) indicate the evidence plot using logFC in Ppert and pnael (b) indicates the eveidence plot using moderated t values in P_{PERT}. Each Pathway is represented by one dot. The pathways at the right of the line (2) are significant after Bonferroni correction of the global p-values, pG. The Pathways at the right of the line (1) are significant after a FDR correction of the global p-values, pG.

Conclusion

Gene set or pathways analysis are more important issue in biological analysis from last decade. Though lot of methods already developed in this case, still more powerful ways need to be developed. In this paper, it was noted that applying moderated t values instead of logFC for calculating probability perturbation gives more significant pathways based on FDR correction. Therefore, it is better to use moderated t values in SPIA where we have less chance to loose the information than logFC.

Bibliography

- SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V.K., MUKHERJEE, S., EBERT, B.L., GILLETTE, M.A., PAULOVICH, A., POMEROY, S.L., GOLUB, T.R., LANDER, E.S. *et al.* (2005), Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43)

- MOOTHA, V.K., LINDGREN, C.M., ERIKSSON, K.F., SUBRAMANIAN, A., SIHAG, S., LEHAR, J., PUIGSERVER, P., CARLSSON, E., RIDDERSTRALE, M., LAURILA, E., *et al.* (2003), PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3).

- EFRON, B., TIBSHIRANI, R. (2007), On testing the significance of set s of genes. *Ann Appl Stat,* 1,p.107-129.

- TOMFOHR, J., LU, J., KEPLER, T.B. (2005), Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6,225.

- LU, Y., LIU, P.Y., XIAO, P., DENG, H.W. (2005), Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, 21(14).

- GOEMAN, J.J., VAN DE GEER, S.A., DE KORT, F., VAN HOUWELINGEN, H.C. (2004), A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1).

- MANSMANN, U., MEISTER, R.(2005), Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med*, 44(3).

- HUMMEL, M., MEISTER, R., MANSMANN, U. (2008), GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics,* 24(1).

- GOEMAN, J.J., MANSMANN, U. (2008), Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24(4).

- TIAN, L., GREENBERG, S.A., KONG, S.W., ALTSCHULER, J., KOHANE, I.S., PARK, P.J. (2005), Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, 102(38).

- CHEN, J.J., LEE , T., DELONGCHAMP, R.R., CHEN, T., TSAI, C.A. (2007), Significance analysis of groups of genes in expression profiling studies. *Bioinformatics*, 23(16).

- DINU, I., POTTER, J.D., MUELLER, T., LIU, Q., ADEWALE, A.J., JHANGRI, G.S., EINECKE, G., FAMULSKI, K.S., HALLORAN, P., YASUI, Y. (2007), Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics,* **8**, 242.

- ADEWALE, A.J., DINU, I., POTTER, J.D., LIU, Q., YASUI, Y. (2008), Pathway analysis of microarray data via regression. *J Comput Biol*, 15(3).

- TSAI, C.A., CHEN, J.J. (2009), Multivariate analysis of variance test for gene set analysis. *Bioinformatics,* 25(7).

- TARCA, A.L., DRAGHICI, S., KHATRI, P., HASSAN, S.S., MITTAL, P., KIM, J.S., KIM, C..J., KUSANOVIC, J.P. and ROMERO, R. (2009), A novel signaling pathway impact analysis. *Bioinformatics,* 25(1).

- SMYTH G..K. (2004), Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3.