ON IMPORTANCE OF TOPOLOGY ON FUNCTIONAL ANNOTATION ON BIOLOGICAL PATHWAYS IN GENE EXPRESSION EXPERIMENTS

Arsen Arakelyan

Abstract: *Nowadays, gene set enrichment analysis (GSEA) is the method of choice for functional annotation in gene expression experiments. However, this approach demonstrates severe inaccuracy, when enrichment of biological pathways is considered. We hypothesize that lack of pathway topology information can be a reason for that. In this paper we evaluated the performance of GSEA and topology-based gene pathway deregulation assessment approach: Pathway Signal Flow (PSF). We demonstrated that PSF outperforms GSEA in pathway deregulation analysis and provided explanations for this observation.*

Keywords: *gene expression analysis, biological pathway, pathway topology, Gene Set Enrichment Analysis, Pathway Signal Flow.*

ACM Classification Keywords: *G.3 Probability and Statistics - Statistical computing, G.3 Probability and Statistics - Statistical software, G.3 Probability and Statistics - Nonparametric statistics*

Introduction

Genome-wide and whole genome transcriptome measurements are usually aimed at the assessment of global changes in gene expression landscape and identification of gene(s) that are differentially expressed under certain condition. Besides the various strategies and statistical approaches developed and used in global gene expression analysis the final output are lists of differentially genes or ranked gene lists. Next, these genes should be functionally annotated in order to better understand the underlying biological processes [Arakelyan et al, 2013].

Nowadays, gene set enrichment analysis (GSEA) is the method of choice for functional annotation in gene expression experiments [Hung, 2013; Subramanian, 2005]. It takes a list of pre-ranked gene list from the user dataset and compares it against known functional gene sets, such as Gene Ontology (GO) categories, microRNA targets, transcription factor binding targets, protein-protein interaction partners, etc. GSEA logic is based on the concept of "small, but concordant" changes [Subramanian, 2005]. While it works well for gene sets joined around molecular function and/or process (such as GO categories, miRNA and TF targets), it has a limited value for gene sets of protein-protein interaction

networks and biological pathways. The reason here is that genes in pathways highly connected to each other with regulatory or physical interactions of opposite directions (activation/inhibition or expression/repression). Thus, overexpression of one gene in pathway may lead to downregulation of others. This scenario has not implemented in GSEA, which searches for gene sets with orchestrated changes of expression towards over- or under-expression [Nersisyan et al., 2016].

Recently, several algorithms were proposed for analysis of pathway involvement based on gene expression data and pathway topology [Draghici et al., 2007; Rahnenführer et al., 2004]. Pathway signal flow (PSF) algorithm developed by us also belongs to this class of algorithms [Nersisyan et al., 2015; Nersisyan et al., 2016]. It allows for the assessment of global pathway involvement, but also identifying specifically deregulated sinks in multibranched pathways. Because PSF takes into account pathway topology and gene expression values; we believe that it should outperform GSEA, when a geneset represents a biological pathway.

In this paper we compared PSF and GSEA in analysis of pathway deregulation in number of diseases characterized by involvement of immune system response.

Data and methods

Data sourses

In this study we analyzed two microarray datasets obtained from Gene Expression Omnibus public repository [Barrett et al., 2011; Barrett and Edgar, 2006] that contain gene expression profiles in disorders characterized by involvement of inflammatory/immune response, such as psoriasis (GSE13355), chronic obstructive pulmonary disease (COPD, GSE42057), multiple sclerosis (MS, GSE13732) and cardioembolic ischemic stroke (IS, GSE58294). From these datasets only samples obtained from untreated patients were used in downstream analyses.

Pathway data used

For comparison purposes we have chosen 24 immune system and cell signaling pathways that were previously shown to be implicated in above mentioned diseased conditions (table 1).

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) was performed using classical algorithm implemented in GSEA Java application by broad institute using default parameters [Subramanian, 2005].

Pathway signal flow analysis

Pathway signal flow (PSF) algorithm evaluates the changes in activity of a given biological pathway depending on the pathway topology and relative gene expression [Arakelyan et al., 2013; Nersisyan et

al., 2015, 2016]. For comparison purposes we have chosen 24 immune system and cell signaling pathways that were previously shown to be implicated in above mentioned diseased conditions (table 1).

Table 1. Immune system and signaling pathways used in analyses.

Pathway names (source KEGG PATHWAY database http://www.kegg.jp/kegg/pathway.html)	
B cell receptor signaling pathway	Hedgehog signaling pathway
Calcium signaling pathway	HIF-1 signaling pathway
Chemokine signaling pathway	Jak-STAT signaling pathway
ErbB signaling pathway	MAPK signaling pathway
Fc epsilon RI signaling pathway	mTOR signaling pathway
Fc gamma R-mediated phagocytosis	Natural killer cell mediated cytotoxicity
FoxO signaling pathway	Notch signaling pathway
NOD-like receptor signaling pathway	PI3K-Akt signaling pathway
TNF signaling pathway	Rap1 signaling pathway
Ras signaling pathway	TGF-beta signaling pathway
RIG-I-like receptor signaling pathway	Toll-like receptor signaling pathway
T cell receptor signaling pathway	VEGF signaling pathway

In order to assess the changes in pathway activities dataset values were converted into natural scale and gene expression values were averaged for each class separately. Then expression fold change (FC) was calculated by division of gene expression in diseased condition to the corresponding values in

healthy controls. FC values for each gene were mapped to corresponding pathway nodes, and were averaged if a node contained more than one gene. After this step an input signal of unity was applied to the pathway source nodes. Then PSF values were calculated at the output nodes. PSF algorithm is calibrated in a way that gene expression of FC=1 at all nodes (normal gene expression) produces PSF=1 values. Values of PSF less than unity refer to pathway de-activation, while PSF>1 indicates pathway activation. Significance of PSF changes at output nodes was assessed by bootstrapping for 200 steps.

Results and discussion

KEGG pathway enrichment in psoriasis

Psoriasis is an immune-mediated, inflammatory and hyperproliferative disease of the skin and joints and conclusive evidence demonstrates that it has a genetic basis [Deng et al., 2016].GSE13355 data series contains log2 transformed gene expression profiles (Affymetrix HG 133 2 Plus platform) for 58 psoriatic patients and 64 normal healthy controls [Nair et al., 2009]. Here we have used GSEA and PSF to evaluate differential gene expression and biological pathway deregulations between skin biopsies taken from healthy patients and at lesion sites from patients. We expected to observe deregulations in immune system related pathways and signaling pathways regulating the immune response.

GSEA analysis identified 5 up-regulated pathways, while PSF detected pathway deregulations in all 24 pathways (Figure 1).

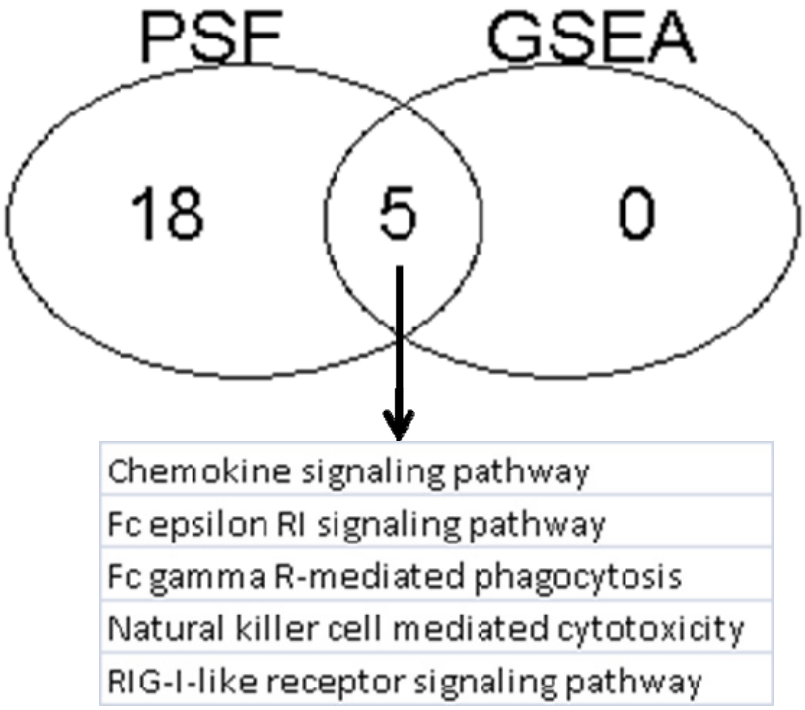

Figure 1. Performance of GSEA and PSF for the psoriasis dataset.

KEGG pathway enrichment in chronic obstructive pulmonary disease

Chronic obstructive pulmonary disease (COPD) is a leading global cause of mortality that is characterized by dysfunction of small airways with cellular inflammation and structural remodeling [Chung and Adcock, 2008]. The GSE42057 dataset contains normalized with robust multi-array averaging method [Irizarry et al., 2003] log2 transformed gene expression profiles of 94 subjects with varying severity of COPD and 45 healthy controls [Bahr et al., 2013]. Gene expression was measured using Affymetrix HG 133 2 Plus platform. GSEA analysis failed to detect pathways associated any pathway associated with inflammation and fibrosis, while PSF correctly identified deregulations in TGF-beta signaling pathway, VEGF signaling pathway as pathways involved in tissue remodeling during COPD and Fc-epsilon RI signaling pathway, T cell signaling pathway, TNF signaling pathway, RIG-I-like receptor signaling pathway and Toll-like receptor signaling pathway as involved in immune response.

KEGG pathway enrichment in multiple sclerosis

Multiple sclerosis (MS) is an organ-specific autoimmune disease caused by an inflammatory demyelinating insult in the central nervous system [Corvol et al., 2008].

We evaluated pathway deregulations base on GSE13732 dataset that contains gene expression profiles measured with Affymetrix HG 133 2 Plus platform in naïve CD4+ T cells of patients with MS before treatment and healthy subjects [Corvol et al., 2008]. GSEA identified TGF-beta signaling pathway as being disturbed in MS patients compared to control. In contrast, PSD detected 21 deregulated pathways (figure 2), from which B cell signaling and T cell signaling are being the most important ones in the pathogenesis of this disorder [Blauthet al., 2015; Holley et al., 2014].

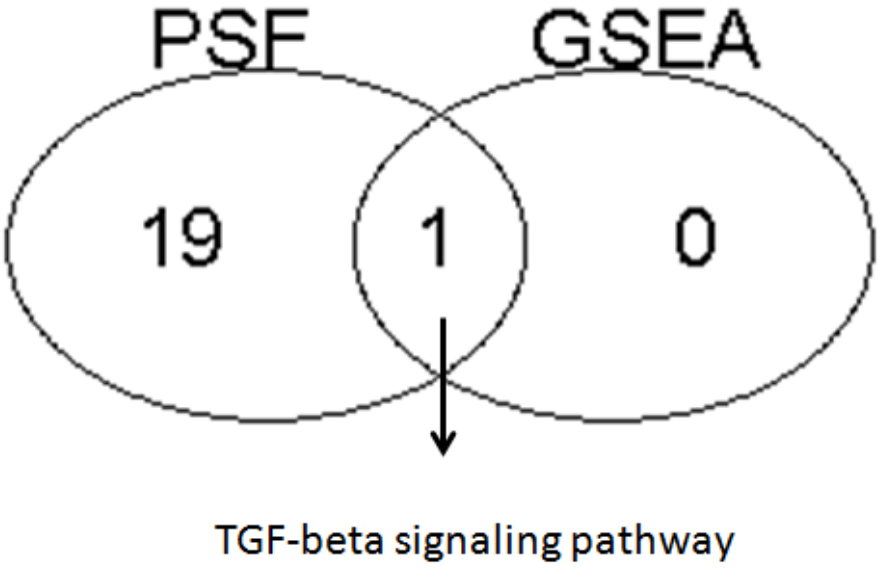

Figure 2. Performance of GSEA and PSF for the multiple sclerosis dataset.

KEGG pathway enrichment in ischemic strokes

Ischemic stroke (IS) accounts for 80% of all strokes, 70% of all acute cerebrovascular diseases, and is the most frequent acute neurological disorder. Its incidence is 14 cases/10000 population/year, and 25% of IS occur in working-age people. Only 1/3 of all stroke patients reach full social and professional reintegration, whereas the remainder die or invalid. The acute phase local and systemic inflammatory response in stroke has shown to play deleterious role along with activated aberrant apoptosis and activation of proinflammatory chemokine signaling [Di Napoli et al., 2006]. In this part of experiments we compared if GSEA and PSF are able to detect the deregulation of inflammation related pathways in stroke patients. We compared gene expression of 24 IS patients (blood collected after 3 hour of insult) with healthy subjects [Stamova et al., 2014]. Again, PSF correctly detects all pathways while GSEA detected only 4 pathways (figure 3).

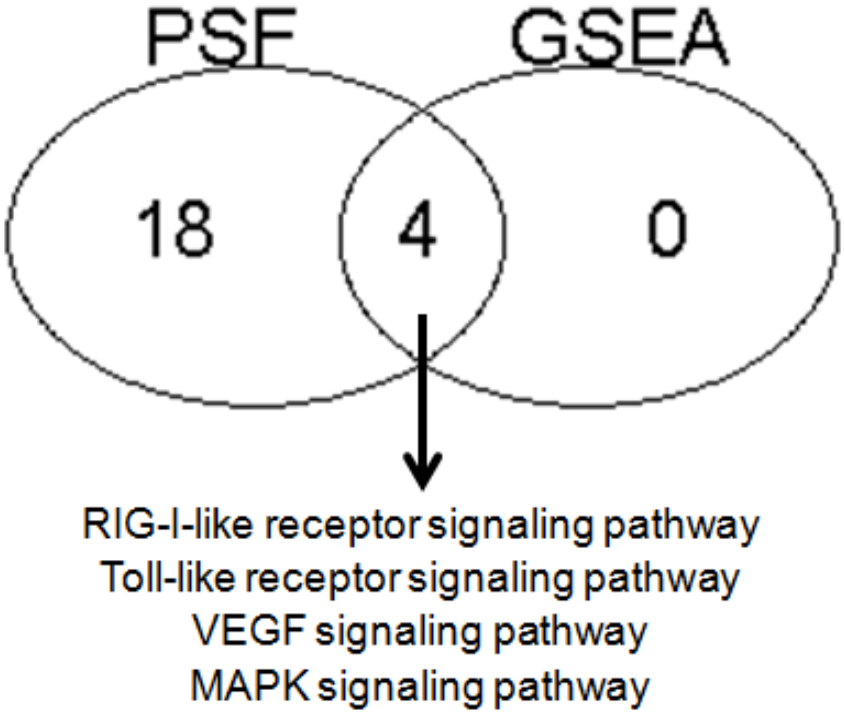

Figure 3. Performance of GSEA and PSF for the ischemic stroke dataset.

Importance of topology for accurate assessment of biological pathway deregulations

The main drawback of the GSEA is its inability to correctly assess the effects of interactions in the pathway. The statistic of GSEA is based on ranking of gene expression while the functional effect of altered gene is neglected. As an example we will analyze TGF signaling pathway deregulation in from multiple sclerosis dataset. The GSEA reported non-significant up-regulation of this pathway ($p = 0.17$, figure 4A). The lack of significance occurred due almost equal number up and down-regulated genes

(figure 4B). However, the closer inspection reveals that the majority of down-regulated genes possess inhibitory action on the on their interacting partners. Downregulation of inhibitory proteins will surely lead to overall up-regulation of this pathway which was actually detected by PSF (table 2). Moreover, PSF allow for evaluation of deregulation at each single branch in multi-branched pathways.

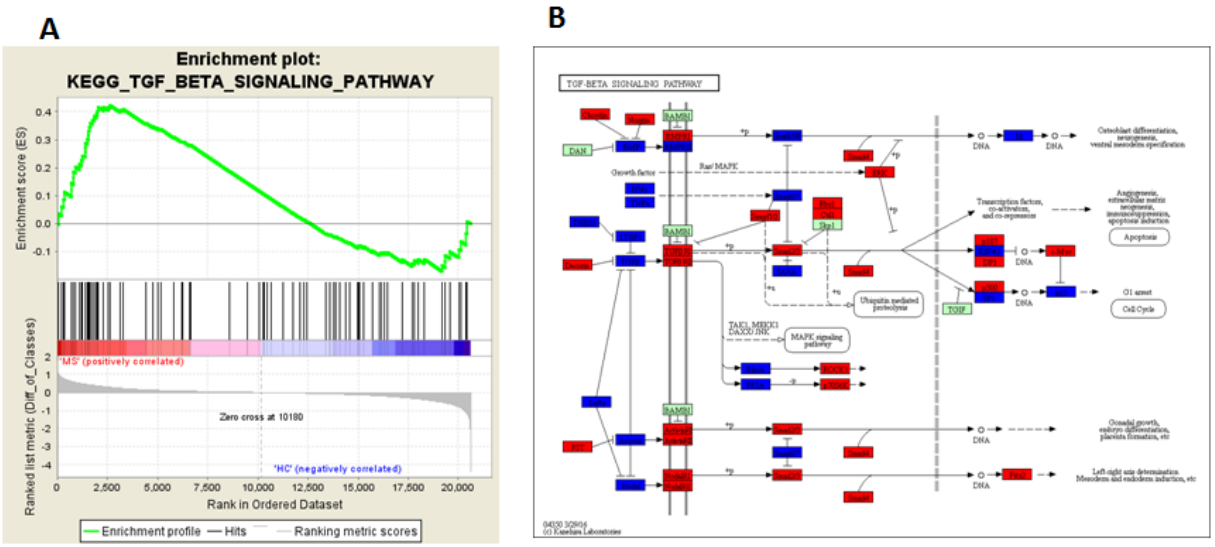

Figure 4. Gene set enrichment statistics score and positions of geneset members on the rank ordered list for TGF-beta signaling pathway (A) and visualization of up- and down-regulated genes in the pathway (B). Note that the most of down-regulated genes inhibit their interaction counterparts.

Table 2. PSF statistic for TGF-beta signaling pathway branches

Pathway	**Branch sink node**	**PSF statistic**	**p value**
TGF-beta signaling pathway	ROCK1	1.361922672	0.045
TGF-beta signaling pathway	RPS6KB1	1.467850941	0.006
TGF-beta signaling pathway	ID1	1.505816732	> 10E-6
TGF-beta signaling pathway	SMAD4	1.590104612	> 10E-6
TGF-beta signaling pathway	SMAD4	1.897146521	> 10E-6
TGF-beta signaling pathway	Cell cycle	0.599811704	> 10E-6
TGF-beta signaling pathway	Apoptosis	1.7290779	> 10E-6

Conclusion

Thus our study demonstrated that different strategy that GSEA should be employed if for analysis of biological pathway deregulations in gene expression studies. While GSEA will work well with sets of independent (i.e. non-interacting) genes, topology should be carefully considered and other methods, such as PSF should be used.

Bibliography

[Arakelyan et al, 2013] A.Arakelyan, L.Aslanyan, A.Boyajyan. High-throughput Gene Expression Analysis Concepts and Applications. In: Ed. iConcept Press Ltd, *Genomics II - Bacteria, Viruses and Metabolic Pathways,* 2013, pp. 71–95.

[Bahr et al, 2013] T.M.Bahr, G.J.Hughes, M.Armstrong, et al. Peripheral Blood Mononuclear Cell Gene Expression in Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory Cell and Molecular Biology*, 2013, *49*(2), 316–323.

[Barrett et al, 2011] T.Barrett, D.B.Troup, S.E.Wilhite, et al. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Research*, 2011, *39*(Database issue), D1005–1010.

[Barrett and Edgar, 2006] T.Barret, R.Edgar. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymology*, 2006, *411, 352-69*.

[Blauth et al, 2015] K.Blauth, G.P. Owens, J.L.Bennett. The Ins and Outs of B Cells in Multiple Sclerosis. *Frontiers in Immunology*, 2015, *6*, 565.

[Chung and Adcock, 2008] K.F.Chung, I.M.Adcock. Multifaceted mechanisms in COPD: inflammation, immunity, and tissue repair and destruction. *The European Respiratory Journal*, 2008, *31*(6), 1334–56.

[Corvol et al, 2008] J.-C.Corvol, D.Pelletier, R.G.Henry, et al. Abrogation of T cell quiescence characterizes patients at high risk for multiple sclerosis after the initial neurological event. *Proceedings of the National Academy of Sciences of the United States of America*, 2008, *105*(33), 11839–44.

[Deng et al, 2016] Y.Deng, C.Chang, Q.Lu. The Inflammatory Response in Psoriasis: a Comprehensive Review. *Clinical Reviews in Allergy & Immunology, 2016*, 1-13.

[Di Napoli et al, 2006] M.Di Napoli, A.Arakelyan, A.Boyajyan, A. (2006). The Acute Phase Inflammatory Response in Stroke: Systemic Inflammation and Neuroinflammation. In: Progress of Inflammation Research. Ed. J.A.Pitzer, New York: Nova Biomedical Books, 2006, 95–145

[Draghici et al, 2007] S.Draghici, P.Khatri, A.L.Tarca, et al. A systems biology approach for pathway level analysis. *Genome Research*, 2007, *17*(10), 1537–45.

[Holley et al, 2007] J.E.Holley, E.Bremer, A.C.Kendall, et al. CD20+inflammatory T-cells are present in blood and brain of multiple sclerosis patients and can be selectively targeted for apoptotic elimination. *Multiple Sclerosis and Related Disorders*, *3*(5), 2007, 650–658.

[Hung 2013] J.H.Hung. Gene set/pathway enrichment analysis. *Methods in Molecular Biology*, 2013, *939*, 201–213.

[Izzary et al, 2013] R.A.Irizarry, B.M.Bolstad, F.Collin. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 2003, *31*(4), e15.

[Nair et al, 2009] R.P.Nair, K.C.Duffin, C.Helms, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature Genetics*, 2009, *41*(2), 199–204.

[Nersisyan et al, 2015] L.Nersisyan, G.Johnson, M.Riel-Mehan, et al. PSFC: a Pathway Signal Flow Calculator App for Cytoscape. *F1000Research,* 2015, *4*.

[Nersisyan et al, 2016] L.Nersisyan, H.Löffler-Wirth, A. Arakelyan, H.Binder. Gene Set- and Pathway-Centered Knowledge Discovery Assigns Transcriptional Activation Patterns in Brain, Blood, and Colon Cancer: *International Journal of Knowledge Discovery in Bioinformatics*, 2016, *4*(2), 46–69.

[Rahnenführer et al, 2004] J.Rahnenführer, F.S.Domingues, J.Maydt, T.Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 2004, *3*, Article16.

[Stamova et al, 2014] B.Stamova, G.C.Jickling, B.P.Ander et al. Gene expression in peripheral immune cells following cardioembolic stroke is sexually dimorphic. *PloS One*, 2014, *9*(7), e102550.

[Subramanian et al, 2005] A.Subramanian, P.Tamayo, V.K.Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 2005, *102*(43), 15545–15550.

Author's Information

Arsen Arakelyan *–Bioinformatics group of the Institute of Molecular Biology NAS RA, Chair of Bioinformatics, Bioengineering and Molecular Biology, Russian-Armenian (Slavonic) University, 7 Hasratyan Str., 0014, Yerevan, Armenia, email: aarakelyan@sci.am*

Major Fields of Scientific Research*: Algorithm development foe gene expression analysis, gene network analysis and modeling*