5034–5044 *Nucleic Acids Research, 2016, Vol. 44, No. 11* *Published online 18 May 2016*
doi: 10.1093/nar/gkw429

SPATIAL: A System-level PAThway Impact AnaLysis approach

Behzad Bokanizad[1,*], Rebecca Tagett[1], Sahar Ansari[1], B. Hoda Helmi[1] and Sorin Draghici[1,2]

[1]Department of Computer Science, Wayne State University, Detroit, MI 48202, USA and [2]Department of Obstetrics & Gynecology, Wayne State University, Detroit, MI 48202, USA

Received May 13, 2015; Revised May 04, 2016; Accepted May 05, 2016

ABSTRACT

The goal of pathway analysis is to identify the pathways that are significantly impacted when a biological system is perturbed, e.g. by a disease or drug. Current methods treat pathways as independent entities. However, many signals are constantly sent from one pathway to another, essentially linking all pathways into a global, system-wide complex. In this work, we propose a set of three pathway analysis methods based on the impact analysis, that performs a system-level analysis by considering all signals between pathways, as well as their overlaps. Briefly, the global system is modeled in two ways: (i) considering the inter-pathway interaction exchange for each individual pathways, and (ii) combining all individual pathways to form a global, system-wide graph. The third analysis method is a hybrid of these two models. The new methods were compared with DAVID, GSEA, GSA, PathNet, Crosstalk and SPIA on 23 GEO data sets involving 19 tissues investigated in 12 conditions. The results show that both the ranking and the *P*-values of the target pathways are substantially improved when the analysis considers the system-wide dependencies and interactions between pathways.

INTRODUCTION

Gene signaling pathways consist of nodes, representing genes or gene products and a set of directed edges between them, representing interactions between genes or gene products. An individual pathway groups genes that work together to drive a certain biological process. The goal of pathway analysis is to identify the pathways that are significantly impacted in a given condition compared to a control (e.g. disease versus healthy, treated versus not-treated, drug A versus drug B, etc.). Pathway analysis requires two types of input: a collection of pathways and a list of genes or gene products that are found to be differentially expressed between the phenotypes compared. Currently, there are several pathway databases providing such collections of pathways for various organisms. Many of these are manually drawn and curated, and updated regularly. Examples include KEGG (1), BioCarta/NCI-PID (2), PANTHER (3) and Reactome (4). The input for pathway analysis is typically the result of high throughput experiments, such as expression data from microarrays, RNA-seq or protein abundance data from mass spectrometry. It consists of thousands of genes/proteins and their corresponding differential expression levels. Depending on the pathway analysis tool used, one or more pathway database(s), and one or more types of biological experiment data may be accepted as input. As output, pathway analysis methods report the pathways that are significantly impacted based on the input data.

Currently, most of the pathway analysis methods consider biological pathways as independent entities. However, the genes in an organism work together as a single system. A perturbation in one pathway impacts other pathways, therefore there are dependencies and interactions between them. Recently, some methods have been developed to detect and correct the cross-talk between pathways (5–7). However, most existing pathway analysis approaches still analyze pathways independently, ignoring interactions and signals that go from one pathway to another, which potentially causes a significant loss of information. Here, we introduce three novel pathway analysis methods that consider all KEGG non-metabolic pathways as a single system, in which the most perturbed pathways are not only identified based on the significance of each pathway individually and independently, but also by considering the impact of other pathways in the global system. We developed and tested three analysis methods using signaling pathway impact analysis (SPIA) (8) method as the base of implementation. These three methods consider the entire system using all interactions between pathways, as well as the network topology.

We test and validate our approaches and compare them with both gene set based methods (9,10) and topology based methods (11). Gene set based methods consider only the set of genes contained in the pathways (its nodes), while topology based methods use both nodes and edges. Since we are

*To whom correspondence should be addressed. Tel: +1 313 577 5070; Fax: +1 313 577 6868; Email: behzad@wayne.edu

using the KEGG pathway database, nodes are genes and edges represent interactions between genes.

Methods for gene set analysis include: *Over-Representation Analysis* (ORA) and *Functional Class Scoring* (FCS). Methods in the ORA category calculate pathway significance by calculating the probability of observing the number of differentially expressed genes in a given pathway by chance alone using the hypergeometric and chi-square statistical tests. Database for annotation, visualization and integrated discovery (DAVID) (12) is one of the ORA based pathway analysis approaches that provides a set of data mining and visualization tools for understanding of biological data. FCS methods consider the position of all genes in the ranked list produced by a selected statistical test for differential expression. Some of FCS methods are, *Gene Set Enrichment Analysis* (GSEA) (13), *Gene Set Analysis* (GSA) (14) and *Pathway Analysis with Down-weighting of Overlapping Genes* (PADOG) (15). The main difference between the ORA and FCS methods is that ORA relies on the selection of a subset of differentially expressed genes, while FCS considers the entire set of genes measured.

Topology-based pathway analysis approaches have been proposed more recently as methods that can integrate both gene set based analysis and signaling interactions between genes, based on the network topology. *Pathway-Express* (16), SPIA (8), *Pathway-Guide* (Advaita Bioinformatics, http://www.advaitabio.com), *TopoGSA* (17) and *Bayesian Pathway Analysis* (BPA) (18) are some of topology-based pathway analysis approaches. Pathway-Express, SPIA and Pathway-Guide capture the impact of the propagation of perturbations from one gene to another, TopoGSA relies on node centrality measures, and BPA, as its name implies, employs Bayesian network.

The idea of analyzing more than one pathway at a time is relatively new and underexplored. Dutta *et al*. (19) introduced an analysis method named 'Pathway analysis using Network information' (PathNet) that uses the idea of *pooled pathways*, or combining of all pathways. They calculate a score for each pathway using a combination of *direct evidence*, which captures the association of each gene with the condition, and *indirect evidence*, that captures the association of each gene's neighbors with the condition, based on connectivity. We have compared the results of (PathNet) with our proposed methods. The first attempt at capturing and analyzing pathway interactions, described as pathway crosstalk, is very recent (5,7). The crosstalk between pathways investigated by Donato *et al*. focuses on the presence of common genes in different pathways. These common genes are often associated with independent biological modules, such as mitochondria, that are important in many different phenomena described by different pathways (e.g. mitochondria play a central role in energy metabolism, Alzheimer's disease, Huntington's disease, etc.).

To the best of our knowledge, the proposed methods in this paper are the first attempts to use the *direct interactions* between pathways as an integral part of the analysis to identify the pathways that are significantly impacted in a given condition.

We compare the results of our methods with one ORA method (DAVID) (12), two FCS methods (GSEA and GSA) (13,14), and three topology-based methods – PathNet (19), Crosstalk (5,7) and SPIA (8). Results are evaluated based on the performance of each method using public data sets with specific target pathways. For example, a data set comparing normal and cancerous colon would have 'colorectal cancer' as the target pathway since we would like any pathway analysis method to identify the colorectal pathway as impacted in this comparison. Similarly, in a study comparing Alzheimer's disease versus healthy samples we would want the Alzheimer's disease pathway from KEGG to be reported as significant. Hence, the Alzheimer's disease pathway will be considered as the target pathway in this condition, etc. This validation method was previously used by PADOG (15). We use here the same set of 23 GEO data sets involving 19 tissues investigated in 12 conditions.

MATERIALS AND METHODS

Map of inter-pathway interactions

At the time of this writing, KEGG included 175 human non-metabolic pathways (signal transduction, biological processes and specific disease pathways). To construct a map of interconnecting KEGG pathways, we used the 'link to another map' and 'link from another map' annotations. In this way, we were able to link one pathway to another through a single gene, which we refer to as an *interface gene*. Interface genes can be found in either source or sink pathways, or both. We define a pathway as source if it influences another pathway using an interface gene, and similarly, a pathway is defined as sink if it receives the influence of a source pathway via an interface gene. In the work by Donato *et al*. (5) the term 'common gene' refers to a gene that is shared between two pathways, while the term 'interface gene' is used for those genes that connect two pathways through biological interactions and signal transduction. We do not connect two pathways that have no interface genes. This method is stringent; it does not include pathways in the full pathway map unless they can be joined using the specific 'link to/from another map' annotation.

Forty-three pathways, shown in Figure 1, were found to have inter-pathway interactions, and are therefore interdependent. Pathways are shown as rounded rectangles around pathway names, and green rectangles represent the genes that interconnect them. The pathways are divided into three groups, which are color-coded based on their relationships with other pathways. The colors of pathway borders indicate their type. Pathways with black borders send direct signals to other pathways but do not receive any such direct signals (sources). Pathways with red borders only receive explicit signals (sinks). Pathways with blue borders both send and receive explicit signals to and from other pathways. These are the *source* and *sink* pathways, respectively. Pathways in the *blue* group are in between sources and sinks in the map. They receive signals from other pathways, and send signals to other pathways.

For instance, the *Notch signaling pathway* is a source pathway, impacting the *MAPK signaling pathway* through the *Notch* interface gene. The *MAPK signaling pathway*, in turn, impacts other pathways, such as the *p53 signaling pathway*, through the *p53* interface gene. The *Apoptosis* pathway is a *sink*, because it doesn't include an interface

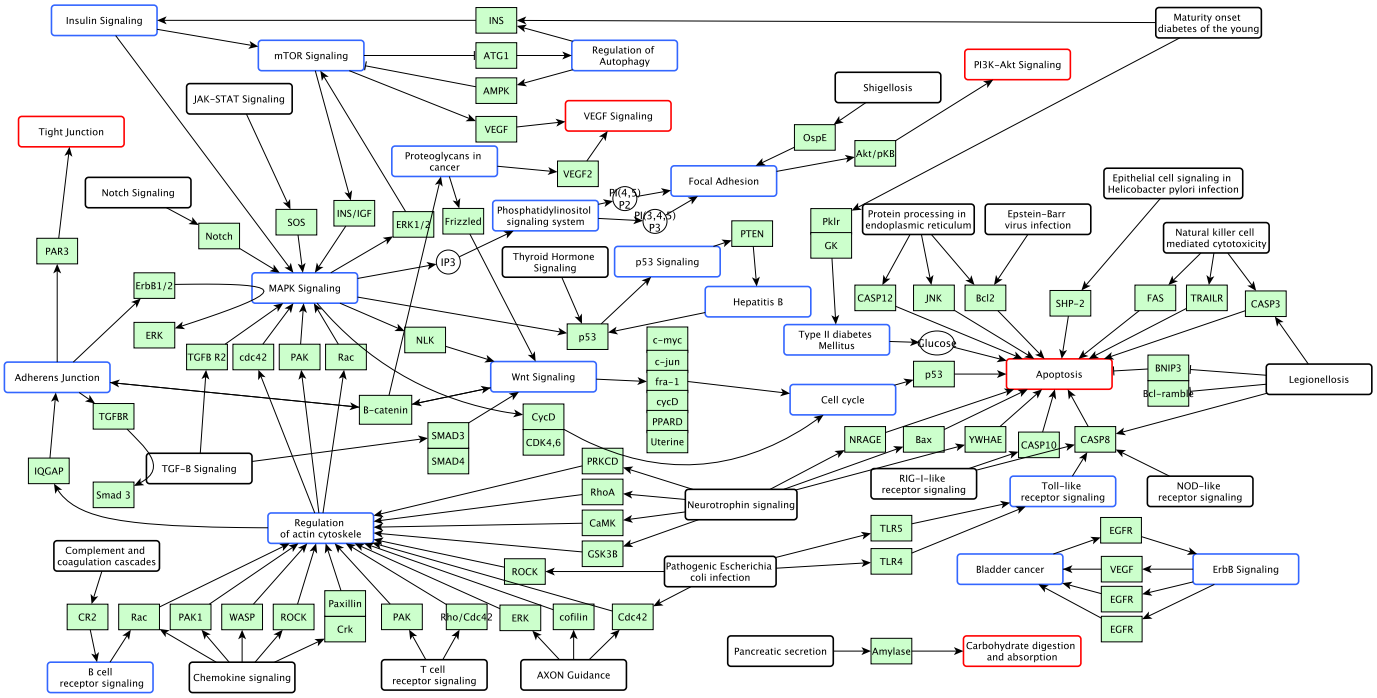

Figure 1. The system-wide map encompassing all KEGG non-metabolic inter-pathway interactions. Pathways are shown as white rectangles. The colors of pathway borders indicate their type. Pathways with black borders send direct signals to other pathways but do not receive any such direct signals (sources). Pathways with red borders only receive explicit signals (sinks). Pathways with blue borders both send and receive explicit signals to and from other pathways. Interface genes are shown as green rectangles. Interface genes can be from either source or sink pathways. Chemical compounds are shown as a circles with the compound names in them.

gene allowing it to directly impact any other pathway, even though the *p53 signaling pathway* and *Cell cycle* are among the pathways shown inside the *Apoptosis* pathway. However, the p53 gene appears again as an interface gene connecting the *Cell Cycle* pathway to *Apoptosis* pathway.

There are several KEGG pathways with interactions in which a gene 'x' activates a gene 'y', in the same pathway, by passing through another pathway. KEGG shows these with a 'link to another map' arrow (since all observed cases are activations), and the signal going from gene 'x' to gene 'y' through pathway 'B'. These interactions are not included in the mathematical model, but are shown in Figure 1 as arching arrows. For example, in the *Adherence Junction* pathway, *ErbB1/2* activates *ERK* by means of the *MAPK signaling pathway*.

In addition to interactions between pathways and genes, KEGG includes some interactions between genes/pathways and DNA or small molecules. These are outside of the scope of this study and are not incorporated in the current analysis, however, they are shown as small circles in Figure 1.

Method 1: System-level PAThway Impact AnaLysis using map (SPATIAL)

The first method, System-level PAThway Impact AnaLysis using the global system map (SPATIAL), combines within-pathway data with the inter-pathway interaction information shown in Figure 1.

For a given pathway P_i, equation 1 expresses a straightforward way of integrating inter-pathway interactions with the results of a topology-based pathway analysis method such as the impact analysis (20). The equation sums the score of each pathway with the score(s) coming from upstream pathways.

$$\begin{aligned} &Pathway\ Impact\ Score\ (P_i) \\ &\quad = Impact\ Score\ from\ topology\ of\ (P_i) \\ &\quad + Impact\ Score(s)\ from\ upstream\ pathways \end{aligned} \quad (1)$$

For the first term in Equation 1, any topology-based technique, such as those discussed in the introduction, can be applied, as long as a score is calculated for each individual pathway. We chose to use SPIA (8), a popular implementation of the impact analysis approach, previously published by our group.

The impact analysis relies on a statistical formulation that combines two probabilities, one from a gene set technique, such as any of those discussed in the introduction section, and another one that accounts for the amount of perturbation on the individual pathway network as a whole, based on the topology of the pathway (20).

For the first probability, we use ORA based on a hypergeometric model. The second probability is calculated based on a perturbation factor (PF) for each gene in the pathway, defined by Equation 2.

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} * \frac{PF(g_j)}{N_{ds}(g_j)} \quad (2)$$

The perturbation factor, PF, for a gene g_i is the expression change (e.g. fold change) of g_i, given by $\Delta E(g_i)$, added to the sum of the n (weighted and normalized) perturbation factors from upstream genes, g_j. $PF(g_j)$ is normalized by the number of genes downstream of g_j, given by $N_{ds}(g_j)$. Thus, the impact of g_j is equally divided between the genes that are directly downstream of it. The βs are weights assigned according to the type and intensity of the reaction. More detail can be found in (8,20).

We modify Equation 1 to incorporate the map of inter-pathway interactions, resulting in Equation 3. The first term in Equation 3 gives the score of a pathway using the topology of that pathway alone; the second term captures the sum of normalized scores of all the upstream pathways. Normalization here means that if a pathway has n downstream pathways, N_{DS}, its impact is divided equally among them. For a specific pathway, the numerator in the first part of Equation 3 sums the absolute values of all of the PFs for all of the genes in that pathway, and the denominator is the normalization factor which adjusts the numerator for technology effects and pathway size. A detailed explanation of the first part of Equation 3 can be found in (20).

$$Acc(P_i) = \frac{\sum_{g \in P_i} |PF(g)|}{|\overline{\Delta E}| * N_{de}(P_i)} + \sum_{j \in U} \frac{Acc(P_j)}{N_{DS_j}} \quad (3)$$

As mentioned earlier, Figure 1 represents all of the inter-pathway interactions that exist in KEGG (1), and includes all of the directed edges between pathways, ignoring the undirected edges as well as pathways with common genes only. The impact of these interactions between pathways is captured by the second term of Equation 3 normalized by the N_{DS} pathways that are downstream from those pathways.

Method 2: Signaling Pathway Impact Analysis - Global Perturbation Factor (SPIA-GPF)

The second method that we propose in this paper is SPIA-GPF, so named because it is inspired originally from SPIA. Whereas SPIA scores pathways as independent entities, here it is adapted to work with a *global* graph of all the pathways. The PF, explained in the previous section, is calculated in the global graph for all of the genes in the unified network.

Unlike the map shown in the Figure 1, the global graph is the union of all nodes and edges in the KEGG non-metabolic pathways. The *ROntoTools* package (21) was used to perform the union of the adjacency matrices for all of them, resulting in a single global adjacency matrix. In order to score the pathways independently, the PFs are extracted from the global graph. $PF_G(g)$ represents the PF of a gene g considering all of its interactions with all other genes in all KEGG non-metabolic pathways.

$$Acc(P_i) = \frac{\sum_{g \in P_i} |PF_G(g)|}{|\overline{\Delta E}| * N_{de}(P_i)} \quad (4)$$

To normalize, the sum of perturbation factors is divided by the absolute mean of expression changes, to remove the impact of technology, and divided by the number of differentially expressed genes, N_{de}, to remove the bias due to the pathway size.

Method 3: System-level PAThway Impact AnaLysis - Global Perturbation Factor (SPATIAL-GPF)

In this section, we present the third method, SPATIAL-GPF, which is a combination of the other two methods. All genes are present in a same physiological system, so while it is relevant to integrate all pathways as a global network, different pathways cross interact, either activating or inhibiting each other. Therefore, it is useful to consider the impact between pathways as well as the global interaction network.

SPATIAL-GPF is designed to capture the information from both methods previously mentioned. First, the PFs of all the genes are calculated using the global graph, then the impact of upstream pathways is applied, using the map of inter-pathway interactions.

$$Acc(P_i) = \frac{\sum_{g \in P_i} |PF_G(g)|}{|\overline{\Delta E}| * N_{de}(P_i)} + \sum_{j \in U} \frac{Acc(P_j)}{N_{DS_j}} \quad (5)$$

Similar to the methods proposed above, we normalize to remove the effect of technology, pathway size and the number of pathways which are downstream of upstream pathways.

RESULTS

To date there is no universally accepted technique for the validation of the results of pathway analysis methods. The assessment of the results of different pathway analysis methods usually involves the selection of a few data sets, and then the interpretation of the results either with the help of a life scientist, or by searching the published literature. This approach is very limited because it can only be applied to a small number of data sets. Furthermore, it is subjective, and may lead to biased results since most of the time the expert who performs the assessment is also a co-author of the paper. Finally, the biological phenomena are so complex that with enough literature search, a large number of pathways can be implicated directly or indirectly in almost any condition. In this work, we follow the validation approach introduced in (15). We use this evaluation approach because it is objective, reproducible, based on multiple data sets, and it does not require an unavoidably biased 'expert' human evaluation of the results (15). This approach requires testing on a large number (at least 10 but preferably more) of different data sets coming from a variety of different conditions, tissues and laboratories. Any number of data sets from any conditions can be used. The only requirement in selecting these data sets is that the condition studied is modeled by a specific pathway in the target database used. For each data set, the pathway corresponding to the phenotype is considered to be the target pathway (e.g. the colorectal cancer pathway will be the target pathway in a colorectal cancer data set). The evaluation focuses on the ability of each method to identify these true positive pathways as significant, and rank them as high as possible.

In this paper, we validated the proposed method using 23 gene expression data sets involving 12 conditions and 19

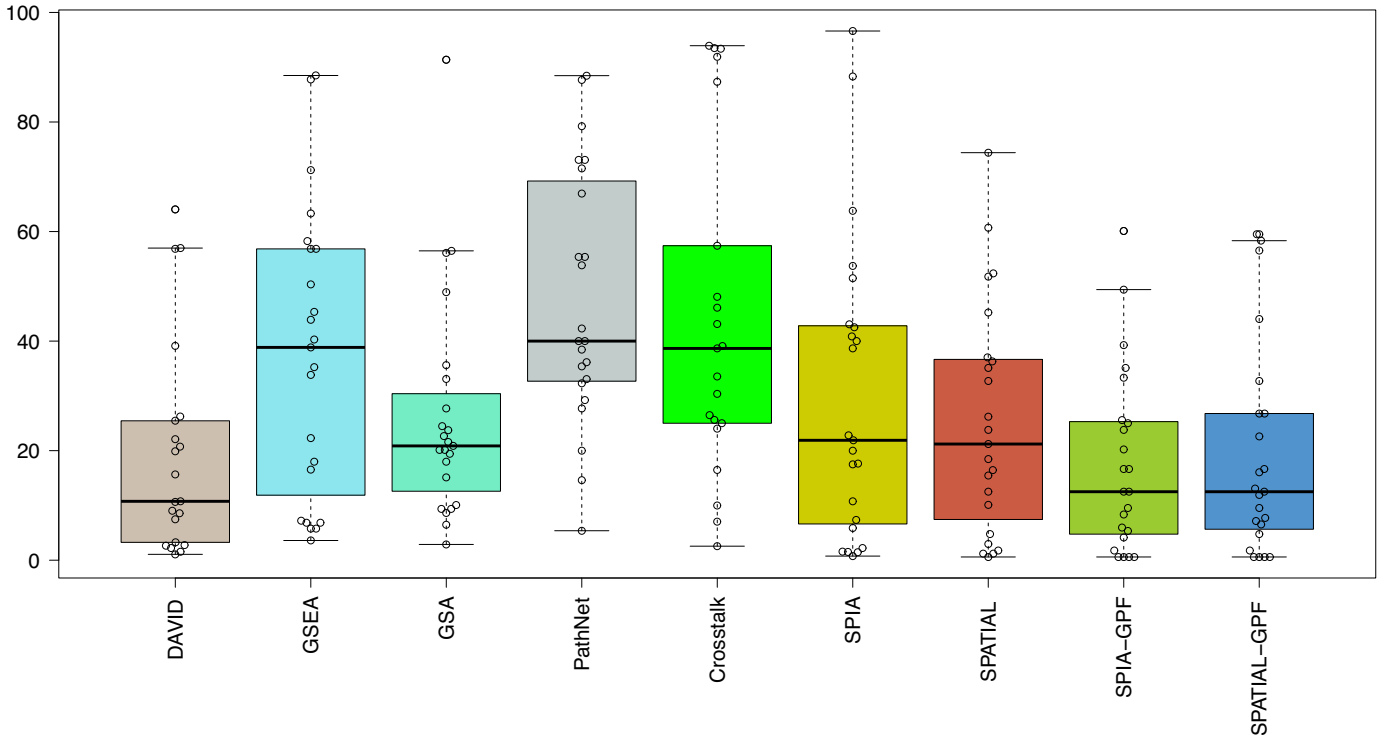

Figure 2. Comparing the *ranks* of the target pathways obtained with DAVID, GSEA, GSA, PathNet, Crosstalk, SPIA, SPATIAL, SPIA-GPF and SPATIAL-GPF on their respective data sets (15). The vertical axis shows the *normalized ranks* of the target pathways on a scale from 0 to 100. Lower values are better.

tissues (see Table 1). We assess the results considering both the rankings and *P*-values of the target pathways associated with the given conditions. Figures 2 and 3 show box plots of normalized ranks and *P*-values of target pathways for all 9 methods tested (the 6 existing methods and the 3 proposed here), using the 23 data sets. The values in Figure 2 are normalized ranks in the range of 0 to 100; here lower values are better. The values in Figure 3 are expressed as $-log(P\text{-}value)$, with higher values representing more significant pathways.

The published methods used for comparison include three gene set-based methods: DAVID (12), (GSEA) (13), (GSA) (14), and three topology-based method – PathNet (19), Crosstalk (5,7) and SPIA (8). Figure 3 shows that GSA was not able to find the target pathway as significant at either 1% or 5% in any of the 23 data sets analyzed. The lowest *P*-value for the target pathways using GSA was 6.95e-02. GSEA was able to report the target pathway as significant in only one case at 1% and three cases at 5%, out of the 23 data sets analyzed. Similarly, DAVID reported five cases at 1% and 8 cases at 5%. PathNet also reported 6 cases at 1% and 7 cases at 5%. Crosstalk was not able to detect any of the target pathways in neither of the threshold levels. SPIA was better, reporting the target as significant in 10 out of the 23 cases at both 1% and 5%. In contrast, all methods performing the analysis at the system-level (SPATIAl, SPIA-GPF, SPATIAL-GPF) yielded better to substantially better results with both SPIA-GPF and SPATIAL-GPF, identifying the targets as significant in 19 out of the 23 cases at the 1% level and in 21 out of the 23 cases at the 5% level of significance.

The results presented so far indicate a clear superiority of the new methods with respect to the gene set methods (GSA and GSEA), but a more limited improvement with respect to the existing SPIA, which is a topological analysis method. Since all three approaches proposed here also use pathway topology, we wanted to investigate further whether the improvements of the novel methods are due mainly to the fact that they use the pathway topology, or to the fact that the new approaches consider the system-level interactions between pathways. In order to do this, we pursued a more detailed comparison between the three novel methods and SPIA in particular.

In addition, SPIA is more appropriate for comparison, because both SPIA as well as all three proposed methods use the impact analysis (20). Thus, if there is an improvement in the results, this can only be attributed to the system-level analysis introduced here, rather than to differences between the underlying analysis approach.

We introduce an *improvement score* that can be used to combine the change in rankings with the change in the *P*-values. To compare the performance of a pair of methods for a given data set, we first consider the rank of the target pathway. If the rank of the target pathway improves in one of the methods, we also consider its *P*-values for each method. Assuming a significance threshold of 1%, we assign a score of +1 if the target pathway changes from *non-significant* to *significant* from the reference to the new method, and a score of −1 if the *P*-value of the target path-

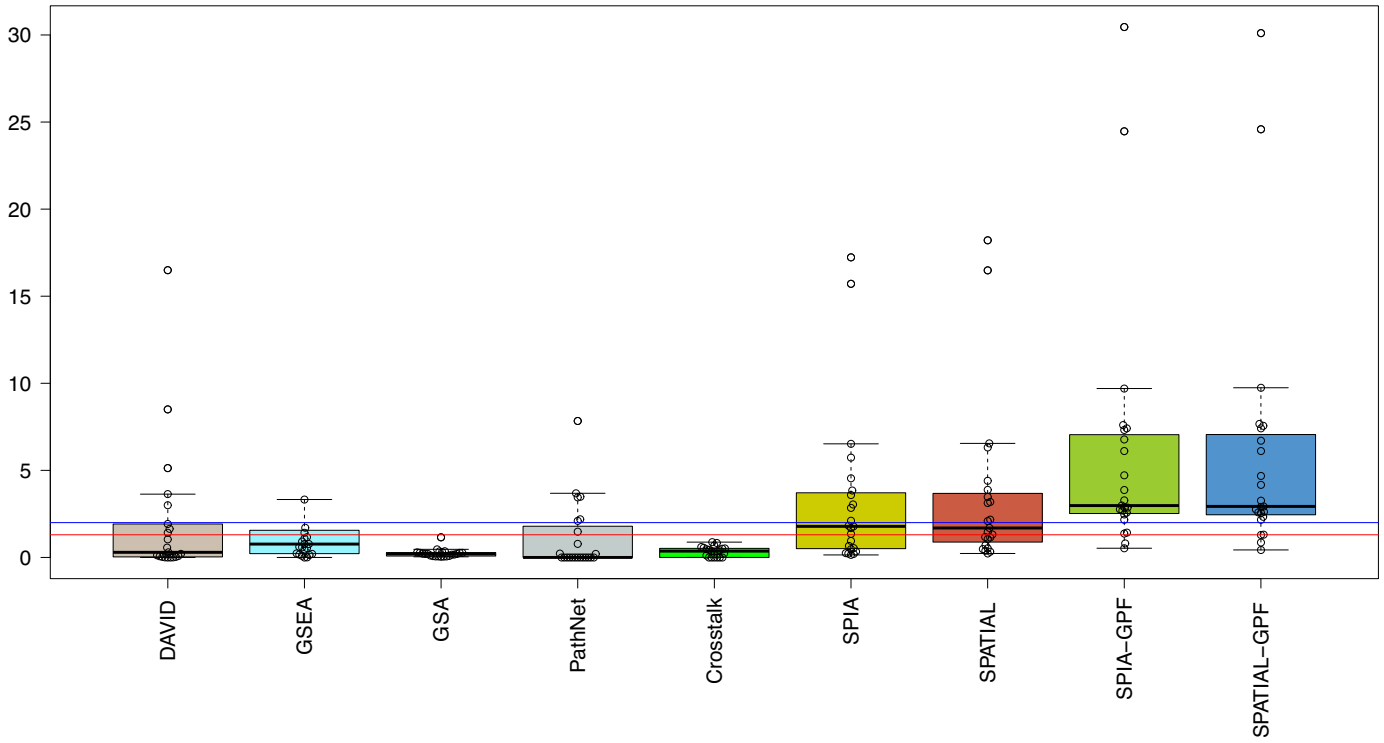

Figure 3. Comparing $-log_{10}(P\text{-}value)$ of the target pathways obtained using DAVID, GSEA, GSA, PathNet, Crosstalk, SPIA, SPATIAL, SPIA-GPF and SPATIAL-GPF on their respective data sets (15). The vertical axis shows the $-log_{10}(P\text{-}values)$ of the target pathways. Higher values show more significance. The red and blue lines represent the 5% and 1% significance levels, respectively.

Table 1. Data sets used for assessing the proposed methods

	Target pathway	KEGG ID	GEO ID	Ref.	Tissue
1	Alzheimer's disease	hsa05010	GSE1297	(22)	Hippocampal CA1
2	Alzheimer's disease	hsa05010	GSE5281	(23)	Brain, Entorhinal cortex
3	Alzheimer's disease	hsa05010	GSE5281	(23)	Brain, hippocampus
4	Alzheimer's disease	hsa05010	GSE5281	(23)	Brain, Primary visual cortex
5	Parkinson's disease	hsa05012	GSE20291	(24)	Postmortem brain putamen
6	Huntington's disease	hsa05016	GSE8762	(25)	Lymphocytes (blood)
7	Colorectal cancer	hsa05210	GSE4107	(26)	Mucosa
8	Colorectal cancer	hsa05210	GSE8671	(27)	Colon
9	Colorectal cancer	hsa05210	GSE9348	(28)	Colon
10	Renal cancer	hsa05211	GSE14762	(29)	Kidney
11	Renal cancer	hsa05211	GSE781	(30)	Kidney
12	Pancreatic cancer	hsa05212	GSE15471	(31)	Pancreas
13	Pancreatic cancer	hsa05212	GSE16515	(32)	Pancreas
14	Glioma	hsa05214	GSE19728	-	Brain
15	Glioma	hsa05214	GSE21354	-	Brain, Spine
16	Prostate cancer	hsa05215	GSE6956	(33)	Prostate
17	Prostate cancer	hsa05215	GSE6956	(33)	Prostate
18	Thyroid cancer	hsa05216	GSE3467	(34)	Thyroid
19	Thyroid cancer	hsa05216	GSE3678	-	Thyroid
20	Acute myeloid leukemia	hsa05221	GSE9476	(35)	Blood, Bone marrow
21	Non-Small Cell Lung Cancer	hsa05223	GSE18842	(36)	Lung
22	Non-Small Cell Lung Cancer	hsa05223	GSE19188	(37)	Lung
23	Dilated cardiomyopathy	hsa05414	GSE3585	(38)	Heart

The 23 data sets used to assess the proposed methods. Each data set comes from a tissue that was affected by a disease/cancer, and the KEGG pathway describing that disease/cancer is assigned for that data set as a target pathway.

way changes from *significant* to *non-significant*. If the rank improves, but does not crosses the significance threshold, we assign a score of +0.5, and if the rank worsens without crossing the significance threshold, we assign a score of −0.5. The improvement score between a pair of methods is calculated as the sum over all tested data sets.

In addition to the proposed improvement score we have used statistical tests including Wilcoxon test and t-tests pairwise between all the methods. In Table 2 we have presented a performance comparison of all the methods using their reported ranks on target pathways. Similarly, we have presented pairwise performance comparison of the *P*-values of the target pathways reported by those methods in Table 3. In both tables, we have highlighted the cells if they are significant at the 5% threshold.

Table 4 shows the comparison between SPIA and the first of our proposed methods, SPATIAL, which considers only inter-pathway interactions. Seventeen of the 23 data sets were better classified using SPATIAL as opposed to SPIA. The result was an overall improvement of 6 for SPATIAL over SPIA, or 26%.

Table 5 shows the comparison between SPIA and the second of our proposed methods, SPIA-GPF, which considers only interactions in the global pathway. A total of 21 out of the 23 data sets were better classified using SPIA-GPF as opposed to SPIA, and just 2 were classified worse. Given the improvement scoring scheme, the result was an overall improvement of 13.5 for SPIA-GPF over SPIA, or 59%.

Table 6 shows the comparison between SPIA and the third of our proposed methods, SPATIAL-GPF, which considers interactions in the global pathway as well as inter-pathway interactions. A total of 22 out of the 23 target pathways were better classified using SPATIAL-GPF as opposed to SPIA, and only 1 was classified worse. The result was an overall improvement score of 15 for SPATIAL over SPIA, equivalent to 65%.

As mentioned earlier, pathways in the KEGG are defined and manually curated by human experts using the existing knowledge resulting from studying biological samples. Some of the interactions that are present in the pathways can be later shown to be wrong (e.g. resulting from errors in experimental measurements) or tissue-specific (e.g. not occurring for all tissues), and are removed in the future releases of the database. New releases of the database includes updates to cover all the new findings. Here, to show that the pathway impact analysis is robust with regard to a small change in the network of pathways, we have modified the input pathways by randomly changing five percent of the nodes and edges present in that pathway. Then, we ran the impact analysis on the new sets of pathways using the same data sets with target pathways. We have compared those results with the previous results by using a t-test with the alternative hypothesis of the true difference in means not being equal to zero. The *P*-values of the ranks of target pathways between the original and randomly changed pathways were 0.96, 0.99 and 0.98 for SPATIAL, SPIA-GPF and SPATIAL-GPF, respectively. Similarly, the *P*-values of the target pathways *P*-values for the comparisons are 0.95, 0.99 and 0.99 for SPATIAL, SPIA-GPF and SPATIAL-GPF, respectively. Based on the results, we cannot reject the null hypothesis. This means that there are no significant changes between the results obtained with the perturbed and original pathways. In other words, the methods proposed are robust with regard to small changes in pathway definitions.

DISCUSSION

Any newly proposed pathway analysis method must be compared to existing methods using real data sets in order to determine its usefulness. In this paper, we compare our three novel methods with six well known analysis methods, three topology-based and three gene set-based.

Traditional validation, using a-posteriori literature-based assessment, is inherently biased since it is performed by a human 'expert', usually a co-author, selecting specific literature as supporting evidence. Furthermore, this type of 'validation' is usually performed on a very small number of data sets. A better assessment approach would eliminate any human bias, could be performed on a large number of data sets and conditions, and could be automated. Such an approach can involve validation using data sets associated with a condition for which there is a specific pathway. For example, if a study compares gene expression from colorectal tumor tissue to normal tissue, the *Colorectal Cancer Pathway* will be the target pathway. Such a pathway is referred to as the 'target' pathway for that data set (15). Any good pathway analysis method should identify this pathway as significant in such a data set.

The target pathway technique has advantages including objectivity, speed of validation and reproducibility. The only disadvantage is that this type of testing focuses on only one pathway for each data set, whereas the behavior of a biological system may be governed by more than one pathway in a given condition. As mentioned, many of these may be relevant to the condition, but for the sake of objectivity, we consider these neither as true positives nor as false negatives addressed because we do not have a priori knowledge of all of the true negatives.

CONCLUSION

In this work, we show that pathway analysis can be significantly improved when inter-pathway interactions are included in the model. This can only be achieved using topology-based methods, allowing the propagation of signal between pathways. We propose three novel approaches: SPATIAL, which considers the effect of signals from another pathways through 'interface genes', SPIA-GPF, which models perturbation through all non-metabolic pathway over an interconnected global map, and SPATIAL-GPF, which combines the first two.

Using 23 data sets, we compare our approach to three gene set based pathway analysis techniques, and three topology based techniques – among them SPIA is the one which was incorporated in our methods. All three of our approaches significantly outperform two of the gene set based methods, GSEA and GSA. DAVID outperforms SPATIAL, but SPIA-GPF and SPATIAL-GPF outperform DAVID.

Table 2. Performance comparison between proposed methods and previously published methods using t-test on the ranks of target pathways

Methods	DAVID	GSEA	GSA	PathNet	Crosstalk	SPIA	SPATIAL	SPIA-GPF	SPATIAL-GPF
DAVID	1	1.279e-02	2.647e-01	7.475e-05	3.120e-03	1.445e-01	3.374e-01	7.640e-01	9.503e-01
GSEA		1	1.109e-01	1.708e-01	4.273e-01	3.503e-01	9.079e-02	4.336e-03	9.635e-03
GSA			1	1.728e-03	2.649e-02	5.944e-01	8.884e-01	1.308e-01	2.259e-01
PathNet				1	6.881e-01	2.286e-02	1.353e-03	1.094e-05	4.18e-05
Crosstalk					1	1.076e-01	2.161e-02	1.211e-03	2.412e-03
SPIA						1	5.195e-01	7.507e-02	1.237e-01
SPATIAL							1	1.824e-01	2.945e-01
SPIA−GPF								1	8.080e-01
SPATIAL−GPF									1

Performance comparison is done pairwise between all the methods discussed in the paper using t-test on the ranks of target pathways reported by each method. The highlighted cells show that the results of the ranking by two methods are statistically different using 5% significance threshold.

Table 3. Performance comparison between proposed methods and previously published methods using t-test on the *P*-values of target pathways

Methods	DAVID	GSEA	GSA	PathNet	Crosstalk	SPIA	SPATIAL	SPIA-GPF	SPATIAL-GPF
DAVID	1	1.984e-01	1.206e-01	2.502e-01	5.375e-01	1.052e-02	1.809e-03	1.454e-04	1.600e-04
GSEA		1	4.000e-04	1.232e-02	2.746e-02	1.172e-01	1.704e-02	4.763e-04	5.577e-04
GSA			1	9.289e-01	2.709e-01	2.132e-08	1.234e-11	2.39e-13	1.5e-13
PathNet				1	4.879e-01	2.200e-04	2.968e-05	2.59e-06	2.795e-06
Crosstalk					1	1.646e-04	8.766e-06	4.062e-07	4.288e-07
SPIA						1	3.486e-01	9.069e-03	1.131e-02
SPATIAL							1	2.958e-02	3.979e-02
SPIA−GPF								1	8.742e-01
SPATIAL−GPF									1

Performance comparison is done pairwise between all the methods discussed in the paper using t-test on the p-values of target pathways reported by each method. The highlighted cells show that the comparison results between two methods are statistically significant using 5% significance threshold.

Table 4. SPIA versus SPATIAL results on target pathways

	Target pathway (TP)	TP Rank SPIA	TP Rank SPATIAL	*P*-value SPIA	*P*-value SPATIAL	Improvement
1	Alzheimer's disease	0.75	0.59	5.918e-18	6.218e-19	0.5
2	Alzheimer's disease	1.45	1.19	1.846e-06	4.814e-07	0.5
3	Alzheimer's disease	2.23	1.78	1.963e-16	3.301e-17	0.5
4	Alzheimer's disease	1.51	1.19	2.994e-07	2.826e-07	0.5
5	Parkinson's disease	63.80	10.11	5.613e-01	6.558e-04	1
6	Huntington's disease	96.61	32.73	6.233e-01	7.077e-02	0.5
7	Colorectal cancer	17.64	16.45	1.637e-02	8.085e-03	1
8	Colorectal cancer	51.47	51.78	2.898e-01	3.297e-01	-0.5
9	Colorectal cancer	38.68	45.23	3.369e-01	4.694e-01	-0.5
10	Renal cancer	17.51	60.71	2.623e-04	2.957e-01	-1
11	Renal cancer	53.73	52.35	7.167e-01	5.919e-01	0.5
12	Pancreatic cancer	22.79	21.21	1.434e-04	1.321e-04	0.5
13	Pancreatic cancer	20.00	18.45	9.075e-04	3.330e-04	0.5
14	Glioma	21.89	15.47	1.427e-03	7.504e-04	0.5
15	Glioma	5.88	2.97	2.834e-05	3.999e-05	0.5
16	Prostate cancer	7.35	23.80	4.458e-02	3.461e-02	-0.5
17	Prostate cancer	10.74	12.50	2.197e-01	8.698e-02	-0.5
18	Thyroid cancer	40.00	35.11	1.118e-01	6.712e-02	0.5
19	Thyroid cancer	42.53	36.30	7.781e-03	4.781e-02	0.5
20	Acute myeloid leukemia	1.57	4.76	1.518e-02	6.765e-03	-0.5
21	Non-Small Cell Lung Cancer	43.06	37.02	2.052e-02	2.038e-02	0.5
22	Non-Small Cell Lung Cancer	88.32	74.40	5.970e-01	4.130e-01	0.5
23	Dilated cardiomyopathy	40.86	26.19	4.631e-01	1.981e-01	0.5
	Sum of improvement scores					6

The normalized ranks and *P*-values produced by SPATIAL and SPIA for the target pathways (TP) in 23 data sets involving 12 conditions. Rankings are normalized on the scale of 1 to 100, and *P*-values are FDR corrected. The scores shows an improvement of 26% (6/23) in SPATIAL compared to the SPIA.

Table 5. SPIA versus SPIA-GPF results on target pathways

	Target pathway (TP)	TP Rank SPIA	TP Rank SPIA-GPF	*P*-value SPIA	*P*-value SPIA-GPF	Improvement
1	Alzheimer's disease	0.75	0.59	5.918e-18	3.503e-31	0.5
2	Alzheimer's disease	1.45	0.59	1.846e-06	4.760e-08	0.5
3	Alzheimer's disease	2.23	1.78	1.963e-16	3.396e-25	0.5
4	Alzheimer's disease	1.51	0.59	2.994e-07	1.998e-10	0.5
5	Parkinson's disease	63.80	39.28	5.613e-01	1.408e-03	1
6	Huntington's disease	96.61	60.11	6.233e-01	3.725e-02	0.5
7	Colorectal cancer	17.64	20.23	1.637e-02	5.322e-04	-0.5
8	Colorectal cancer	51.47	9.52	2.898e-01	6.873e-03	1
9	Colorectal cancer	38.68	33.33	3.369e-01	2.958e-01	0.5
10	Renal cancer	17.51	35.11	2.623e-04	1.051e-03	-0.5
11	Renal cancer	53.73	23.80	7.167e-01	4.215e-02	0.5
12	Pancreatic cancer	22.79	12.50	1.434e-04	3.856e-08	0.5
13	Pancreatic cancer	20.00	8.33	9.075e-04	1.668e-07	0.5
14	Glioma	21.89	16.66	1.427e-03	1.919e-05	0.5
15	Glioma	5.88	4.16	2.834e-05	2.512e-08	0.5
16	Prostate cancer	7.35	5.95	4.458e-02	1.555e-01	0.5
17	Prostate cancer	10.74	5.35	2.197e-01	2.049e-03	1
18	Thyroid cancer	40.00	25.59	1.118e-01	1.369e-03	1
19	Thyroid cancer	42.53	12.50	7.781e-03	2.663e-03	0.5
20	Acute myeloid leukemia	1.57	0.59	1.518e-02	7.838e-07	1
21	Non-Small Cell Lung Cancer	43.06	25.00	2.052e-02	3.388e-03	1
22	Non-Small Cell Lung Cancer	88.32	49.40	5.970e-01	1.728e-03	1
23	Dilated cardiomyopathy	40.86	16.66	4.631e-01	1.341e-04	1
	Sum of improvement scores					13.5

The normalized ranks and *P*-values produced by SPIA-GPF and SPIA for the target pathways (TP) in 23 data sets involving 12 conditions. Rankings are normalized on a scale of 1 to 100, and *P*-values are FDR corrected. The scores shows an improvement of 59% (13.5/23) in SPIA-GPF compared to SPIA.

Table 6. SPIA versus SPATIAL-GPF results on target pathways

	Target pathway (TP)	TP Rank SPIA	TP Rank SPATIAL-GPF	*P*-value SPIA	*P*-value SPATIAL-GPF	Improvement
1	Alzheimer's disease	0.75	0.59	5.918e-18	7.803e-31	0.5
2	Alzheimer's disease	1.45	0.59	1.846e-06	2.171e-08	0.5
3	Alzheimer's disease	2.23	1.78	1.963e-16	2.594e-25	0.5
4	Alzheimer's disease	1.51	0.59	2.994e-07	1.804e-10	0.5
5	Parkinson's disease	63.80	44.04	5.613e-01	2.378e-03	1
6	Huntington's disease	96.61	56.54	6.233e-01	4.926e-02	0.5
7	Colorectal cancer	17.64	16.04	1.637e-02	5.449e-04	1
8	Colorectal cancer	51.47	9.52	2.898e-01	6.873e-03	1
9	Colorectal cancer	38.68	32.73	3.369e-01	3.702e-01	0.5
10	Renal cancer	17.51	58.33	2.623e-04	1.260e-03	-0.5
11	Renal cancer	53.73	26.78	7.167e-01	5.177e-02	0.5
12	Pancreatic cancer	22.79	12.50	1.434e-04	3.856e-08	0.5
13	Pancreatic cancer	20.00	7.73	9.075e-04	1.980e-07	0.5
14	Glioma	21.89	16.66	1.427e-03	2.068e-05	0.5
15	Glioma	5.88	4.76	2.834e-05	2.755e-08	0.5
16	Prostate cancer	7.35	6.54	4.458e-02	1.373e-01	0.5
17	Prostate cancer	10.74	7.14	2.197e-01	2.747e-03	1
18	Thyroid cancer	40.00	22.61	1.118e-01	1.172e-03	1
19	Thyroid cancer	42.53	13.09	7.781e-03	2.602e-03	0.5
20	Acute myeloid leukemia	1.57	0.59	1.518e-02	7.838e-07	1
21	Non-Small Cell Lung Cancer	43.06	26.78	2.052e-02	4.594e-03	1
22	Non-Small Cell Lung Cancer	88.32	59.52	5.970e-01	1.757e-03	1
23	Dilated cardiomyopathy	40.86	11.90	4.631e-01	6.758e-05	1
	Sum of improvement scores					15

The normalized ranks and *P*-values produced by SPATIAL-GPF and SPIA for the target pathways (TP) in 23 data sets involving 12 conditions. Rankings are normalized to the scale of 1 to 100, and *P*-values are FDR corrected. The scores shows an improvement of 65% (15/23) in SPATIAL-GPF compared to SPIA.

ACKNOWLEDGEMENTS

This research was supported in part by the following grants: NIH R01 DK089167, R42 GM087013 and NSF DBI-0965741, and by the Robert J. Sokol, MD Endowment in Systems Biology. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

FUNDING

National Institutes of Health (NIH) [R01 DK089167, R42 GM087013, in part]; National Science Foundation (NSF) [DBI-0965741, in part]; Robert J. Sokol, MD Endowment in Systems Biology (in part). Funding for open access charge: NIH [R01 DK089167, R42 GM087013, in part]; NSF [DBI-0965741, in part]; Robert J. Sokol, MD Endowment in Systems Biology (in part).
Conflict of interest statement. None declared.

REFERENCES

1. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.
2. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37**(Suppl. 1), D674–D679.
3. Mi,H., Lazareva-Ulitsky,B., Loo,R., Kejariwal,A., Vandergriff,J., Rabkin,S., Guo,N., Muruganujan,A., Doremieux,O., Campbell,M.J. *et al.* (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, **33**(Suppl. 1), D284–D288.
4. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G., Wu,G., Matthews,L. *et al.* (2005) REACTOME: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
5. Donato,M. and Drăghici,S. (2010) Signaling pathways coupling phenomena. In: *Neural Networks (IJCNN), The 2010 International Joint Conference on Barcelona*. IEEE, Spain, pp. 1–6.
6. Donato,M., Zhu,Z., Tomoiaga,A., Westfall,P., Romero,R. and Drăghici,S. (2012) A method for analysis and correction of cross-talk effects in pathway analysis. In: *Neural Networks (IJCNN), The 2012 International Joint Conference on Brisbane*. QLD IEEE, Australia, pp. 1–7.
7. Donato,M., Xu,Z., Tomoiaga,A., Granneman,J.G., MacKenzie,R.G., Bao,R., Than,N.G., Westfall,P.H., Romero,R. and Drăghici,S. (2013) Analysis and correction of crosstalk effects in pathway analysis. *Genome Res.*, **23**, 1885–1893.
8. Tarca,A.L., Drghici,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**, 75–82.
9. Khatri,P. and Drăghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
10. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375.
11. Mitrea,C., Taghavi,Z., Bokanizad,B., Hanoudi,S., Tagett,R., Donato,M., Voichiţa,C. and Drăghici,S. (2013) Methods and approaches in the topology-based analysis of biological pathways. *Front. Physiol.*, **4**, 278.
12. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.
13. 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**, 15545–15550.
14. Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.
15. Tarca,A.L., Drghici,S., Bhatti,G. and Romero,R. (2012) Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, **13**, 136.
16. Khatri,P., Drăghici,S., Tarca,A.L., Hassan,S.S. and Romero,R. (2007) A system biology approach for the steady-state analysis of gene signaling networks. In: *CIARP'07 Proceedings of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications Valparaiso*, ACM, Chile, pp. 32–41.
17. Glaab,E., Baudot,A., Krasnogor,N. and Valencia,A. (2010) TopoGSA: network topological gene set analysis. *Bioinformatics*, **26**, 1271–1272.
18. Isci,S., Ozturk,C., Jones,J. and Otu,H.H. (2011) Pathway analysis of high-throughput biological data within a Bayesian network framework. *Bioinformatics*, **27**, 1667–1674.
19. Dutta,B., Wallqvist,A. and Reifman,J. (2012) PathNet: A tool for pathway analysis using topological information. *Source Code Biol. Med.*, **7**, 10.
20. Drăghici,S., Khatri,P., Tarca,A.L., Amin,K., Done,A., Voichiţa,C., Georgescu,C. and Romero,R. (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545.
21. Voichiţa,C., Donato,M. and Drăghici,S. (2012) Incorporating gene significance in the impact analysis of signaling pathways. In: *Proceedings of the International Conference on Machine Learning Applications (ICMLA)*. pp. 126–131.
22. Blalock,E.M., Geddes,J.W., Chen,K.C., Porter,N.M., Markesbery,W.R. and Landfield,P.W. (2004) Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 2173–2178.
23. Liang,W.S., Dunckley,T., Beach,T.G., Grover,A., Mastroeni,D., Walker,D.G., Caselli,R.J., Kukull,W.A., McKeel,D., Morris,J.C. *et al.* (2007) Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics*, **28**, 311–322.
24. Zhang,Y., James,M., Middleton,F.A. and Davis,R.L. (2005) Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. *Am. J. Med. Genet. B*, **137**, 5–16.
25. Runne,H., Kuhn,A., Wild,E.J., Pratyaksha,W., Kristiansen,M., Isaacs,J.D., Régulier,E., Delorenzi,M., Tabrizi,S.J. and Luthi-Carter,R. (2007) Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14424–14429.
26. Hong,Y., Ho,K.S., Eu,K.W. and Cheah,P.Y. (2007) A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin. Cancer Res.*, **13**, 1107–1114.
27. Sabates-Bellver,J., Van der Flier,L.G., de Palo,M., Cattaneo,E., Maake,C., Rehrauer,H., Laczko,E., Kurowski,M.A., Bujnicki,J.M., Menigatti,M. *et al.* (2007) Transcriptome profile of human colorectal adenomas. *Mol. Cancer Res.*, **5**, 1263–1275.
28. Hong,Y., Downey,T., Eu,K.W., Koh,P.K. and Cheah,P.Y. (2010) A 'metastasis-prone'signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. *Clin. Exp. Metastasis*, **27**, 83–90.
29. Wang,Y., Roche,O., Yan,M.S., Finak,G., Evans,A.J., Metcalf,J.L., Hast,B.E., Hanna,S.C., Wondergem,B., Furge,K.A. *et al.* (2009) Regulation of endocytosis via the oxygen-sensing pathway. *Nat. Med.*, **15**, 319–324.
30. Lenburg,M.E., Liou,L.S., Gerry,N.P., Frampton,G.M., Cohen,H.T. and Christman,M.F. (2003) Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer*, **3**, 31.
31. Badea,L., Herlea,V., Dima,S.O., Dumitrascu,T. and Popescu,I. (2008) Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology*, **55**, 2015–2026.

32. Pei,H., Li,L., Fridley,B.L., Jenkins,G.D., Kalari,K.R., Lingle,W., Petersen,G., Lou,Z. and Wang,L. (2009) FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. *Cancer Cell*, **16**, 259–266.
33. Wallace,T.A., Prueitt,R.L., Yi,M., Howe,T.M., Gillespie,J.W., Yfantis,H.G., Stephens,R.M., Caporaso,N.E., Loffredo,C.A. and Ambs,S. (2008) Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res.*, **68**, 927–936.
34. He,H., Jazdzewski,K., Li,W., Liyanarachchi,S., Nagy,R., Volinia,S., Calin,G.A., Liu,C.-g., Franssila,K., Suster,S. *et al.* (2005) The role of microRNA genes in papillary thyroid carcinoma. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 19075–19080.
35. Stirewalt,D.L., Meshinchi,S., Kopecky,K.J., Fan,W., Pogosova-Agadjanyan,E.L., Engel,J.H., Cronk,M.R., Dorcy,K.S., McQuary,A.R., Hockenbery,D. *et al.* (2008) Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer*, **47**, 8–20.
36. Sanchez-Palencia,A., Gomez-Morales,M., Gomez-Capilla,J.A., Pedraza,V., Boyero,L., Rosell,R. and Fárez-Vidal,M.E. (2011) Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int. J. Cancer*, **129**, 355–364.
37. Hou,J., Aerts,J., Den Hamer,B., Van Ijcken,W., Den Bakker,M., Riegman,P., van der Leest,C., van der Spek,P., Foekens,J.A., Hoogsteden,H.C., Grosveld,F. *et al.* (2010) Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One*, **5**, e10312.
38. Barth,A.S., Kuner,R., Buness,A., Ruschhaupt,M., Merk,S., Zwermann,L., Kääb,S., Kreuzer,E., Steinbeck,G., Mansmann,U. *et al.* (2006) Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *J. Am. Coll. Cardiol.*, **48**, 1610–1617.