A SIGNALING PATHWAY ANALYSIS METHOD BASED ON INFORMATION DIVERGENCE

Hang Wei [1], *Hao-Ran Zheng* [1,2]

[1]School of Computer Science and Technology and [2]Department of Systems Biology,
University of Science and Technology of China, Hefei 230026, China
E-mail: hangwei@mail.ustc.edu.cn

Keywords: Pathway analysis, information divergence, microarray expression data, difference analysis

Abstract

Abnormal regulation of signaling pathways is the key factor causing disease. For better understanding disease mechanisms, many methods have been proposed to identify the significantly differential pathways between diseases and normal individuals via microarray gene expression datasets. Unlike previous common analysis processes, which is focused on merging gene difference into difference of pathway indirectly. In this paper, the idea of information divergence is introduced and a novel signaling pathway analysis method from a holistic view is presented to improve the detection results. We identify significantly differential pathways directly via computing the KL divergence between real and simulated probability distributions of gene-gene regulatory ability. We test our method on four human microarray expression datasets. The results illustrate that the capability of our approach in detecting significantly differential pathways between two sample groups is superior to other three classical pathway analysis methods.

1 Introduction

Signaling pathway is a major biological function of organism, and checking the activity of signaling pathway will reflect the current physiological state of cell. However, a large number of biological experiments need to be implemented to determine the activity of signaling pathway via wet experiment method. With the popularity of microarray chip technology, a growing number of researchers are concentrated on studying signaling pathway using bioinformatics. In order to predict possible disease marker and help to understand disease mechanisms, the microarray gene expression data was used to identify significantly differential pathway between normal and cancer samples.

The objective of pathway analysis is to find significantly differential pathways between two given conditions [1]. In the past few decades, pathway analysis has experienced three generations. The first two generation methods referred to as gene set difference analysis [2]. These methods are based on considering pathways as simple gene lists. The first generation approaches are known as Over-Representation Analysis (ORA), such as GOToolBox [3], GOEAST [4]. The differential pathway is inferred via estimating the chance of observing a given number of genes from a particular pathway among the selected differentially expressed genes. The second generation methods are called Functional Class Scoring (FCS). Subramanian et al. proposed the earliest FCS method named GSEA [5]. It selects a statistic to compare groups of samples, and then rank the entire list of genes according to the value of the statistic, finally predict significantly differential pathways by permutation test. As a result, most methods were improved on GSEA, such as ssGSEA [6], GANPA [7].

Although ORA and FCS are capable of providing list of pathway with differentially expressed genes, they are some issues still need to overcome. Firstly, ORA discard gene expression information, and only consider number of differential expressed genes. Moreover, Both ORA and FCS identify pathway via differentially expressed (DE) genes, which consider pathways as simple gene lists, ignoring changes of relationship among genes [2]. To solve those problems, the third generation approaches considered the integration of gene expression and pathway topology. Some of these approaches identify pathway based on differentially co-expressed (DC) genes, such as GSCA [8], GSNCA [9], EDDY [10]. Recently, Ma et al. proposed DRAGEN [11] method, it detects pathway by differentially regulated (DR) genes quantitatively. These third generation methods considered the relationships between genes, and have been widely used. Even though, there still have two issues need to overcome. First of all, the expression of genes is a complicated process regulated by several factors. Gene expression data will reflect relationship between genes at a certain extent, but they will not quantify the association between genes accurately completely. One study reveals that sometimes the relationship between genes computed from gene expression data and the actual biological regulation of genes are inconsistent [12]. Hence most existing approaches, which quantify degree of interaction among genes using gene expression alone, may bring interference information. Secondly, it is general for current methods to merge difference of DE, DC or DR genes independently by simple sum, mean and K-S methods. The above whole workflow pays more attention to the difference of pathway components than systemic difference of pathway. So, these methods may overlook the overall change of pathway.

With the above consideration, we propose a new approach. It is based on one biological hypothesis that biological systems can show highly diverse activity patterns across specific molecular contexts [13]. Our method reflects pathway activity patterns by the probability distribution of gene-gene regulatory capacity. A linear regression model was adopted to simulate the gene-gene regulatory ability using gene expression data. And also an improved directed random walk algorithm on the prior pathway topology was applied to quantify the actual gene-gene regulatory capacity. Then, the concept of information divergence [14] in information theory was introduced to quantify the difference between real and simulated probability distributions of regulatory capacity. Finally, we got pathway distance between two phenotypes according to difference of two KL divergences, and detect final significantly differential pathways by permutation test. During the test, four human microarray expression datasets were implemented and the experiment results indicated that out method achieves better results than previous methods.

The paper is organized as follows: Section 2 describes the detail of our approach. Section 3 indicates the experimental results on different microarray gene expression datasets. At last, Section 4 presents our conclusion.

2 Methods

The following five steps are the main process of our method.

2.1 Data pre-processing

All data manipulations were performed using R 3.1.1. We downloaded Affymetrix CEL files from NCBI GEO [15] and preprocessed all the microarray datasets using RMA method from Bioconductor package. The work described here used human pathways from KEGG [16]. We adopted parseKGML function from KEGGgraph package and remained pathways which contain at least one of the following regulations: inhibition ubiquination, activation, expression, inhibition, activation phosphorylation, inhibition phosphorylation, dephosphorylation inhibition, activation dephosphorylation, inhibition ubiquitination, repression. Finally, we got 166 non-metabolic KEGG pathways.

2.2 Quantization of real gene-gene regulatory ability by DRW

In this study a target signaling pathway is represented as a directed graph, where the nodes represent genes and directed edges represent regulation between genes. *g.start* and *g.end* represent transcription factor and target gene respectively. *I = (g.start,g.end)* is a matrix of regulatory type, and its elements are set according to regulatory type of gene pair. *I(g.start,g.end)* = 1 if *g.start* activates or expresses *g.end*, *I(g.start,g.end)* = –1 if *g.start* inhibits or represses *g.end*. *I(g.start,g.end)* = 0 if *g.start* and *g.end* have no relationship.

Random walk is a classical method for measuring correlation between nodes. It has been applied to pathway analysis [17] and disease classification [18] studies successfully. We suppose that association between *g.start* and *g.end* is closer, thus *g.start* has stronger regulatory ability on *g.end*. Based on this assumption, we introduce directed random walk (DRW) with restart probability method to quantify the real biological regulatory ability of gene pairs, and set restart probability *p*=0.7 initially. We improve traditional DRW method considering the characteristic of biological pathway data. The random walk on the pathway graph and the PageRank algorithm used by the Google search engine are similar. However, considering the importance of transcription factor, the direction of the random walk is set to be opposite. Moreover, in order to distinguish the direct and indirect regulation. We add one step, multiplying original *u* by *exp(-num)* during the iteration loop, where *num* represents iteration time. As the increasing of *num*, the distance between transcription factor and target gene is farther, and the ability of transcription factor regulate target gene is decaying. Thus, these improvement plans can ensure that if genes are close, then the ability of regulation is strong and vice versa. The specific algorithm of DRW with restart probability is illustrated in Table 1.

Table 1. Algorithm of DRW with restart probability

Directed Random Walk with restart probability

```
Input： matrix D for target gene expression data D=(gene,sample), matrix G for one pathway G=(g.start,g.end),with regulation type I=(g.start,g.end), restart probability p.
Output：matrix R.real for gene-gene regulatory degree in pathway
Map pathway G onto rownames of D,and reverse direction of edge regulation in the pathway G, get G':=(g.end,g.start);
for i <- 1 to length |g.end| do {
    v := vector of length |g.end| set v[i] to 1,otherwise 0;
    u := v; u.old := v; num :=0;
    while( TRUE ) do {
        u := (exp(-num))*(1-p)*G'*u.old+p*v
        if ( sum(|u-u.old|) ) < 1E-10)
            break;
        u.old := u;
        num++;
    }
    R_real [,i] := u*I(g_start,g_end);
}
return R.real;
```

2.3 Simulation of gene-gene regulatory ability by gene expression data

Due to gene expression data can reflect ability of genes' regulation at a certain extent, and detect changes in regulatory relationships can discover pathways in response to perturbed phenotypes. Using the idea of DRAGEN for reference, target gene is regulated by transcription factor, and the linear regression model demonstrates reasonably good power in detecting differentially regulated patterns [11].Therefore, we build linear regression model for normal and disease sample to explain the expression levels of the target gene by those of the transcription factors as Eq. (1), where the subscript *i* indexes the normal (*i*=0) or the disease (*i*=1) phenotype. For phenotype *i*, X_i and Y_i represent the expression level of transcription factor and target gene, respectively, α_i and β_i are the regression intercept and slope, ε_i is a zero mean Gaussian noise. In fact, the regression coefficients β_i denote the capacity of regulation between existing gene pairs in pathway. Eventually, for a candidate pathway, we will obtain regulatory capacities of gene pairs ***R.simu_normal*** and ***R.simu_disease*** for two different phenotypes.

$$Y_i = \alpha_i + \beta_i * X_i + \varepsilon_i \tag{1}$$

2.4 Computation of KL divergence

For a specific pathway, *P* denotes real probability distributions, which is gained by DRW. ***Q_normal*** and ***Q_disease*** represent simulated probability distributions for normal and disease sample respectively, and they are computed by liner regression model using gene expression data. The KL divergence between two different distributions can be calculated as Eq. (2) and Eq. (3), where *P(i)*, ***Q_normal*** and ***Q_disease*** can obtain by traditional normalization method as Eq. (4), Eq. (5) and Eq. (6), where *edge* represents existing gene pairs in one candidate pathway.

$$KL\_normal = D(P \| Q\_normal) = \sum P(i)\log(P(i)/Q\_normal(i)) \tag{2}$$

$$KL\_disease = D(P \| Q\_disease) = \sum P(i)\log(P(i)/Q\_disease(i)) \tag{3}$$

$$P(i) = \frac{R.real(i)}{\sum_{j\in edge} R.real(j)} \tag{4}$$

$$Q\_normal(i) = \frac{R.simu\_normal(i)}{\sum_{j\in edge} R.simu\_normal(j)} \tag{5}$$

$$Q\_disease(i) = \frac{R.simu\_disease(i)}{\sum_{j\in edge} R.simu\_disease(j)} \tag{6}$$

2.5 Evaluation of statistical significance

With KL divergence for regulatory capacity probability distributions obtained, we calculate the difference between *KL_normal* and *KL_disease* by absolute difference as Eq. (7), and evaluate the statistical significance using permutation test. In order to retain the gene-gene correlations, we repeat the random shuffling of phenotype labels, and compute *KL_diff* using the permutation data. At last, we obtain *p.value* as Eq. (8), where *I()* is an indicator function, and *n* represents permutation time.

$$KL\_diff = | KL\_normal - KL\_disease | \tag{7}$$

$$p.value = \frac{\sum_{j=1...n} I(KL\_diff_j \geq KL\_diff)}{n} \tag{8}$$

In the following section, we tested our method on four human microarray data, and the results performed by our method are more effective than other three existing approaches.

3 Results

We used four publicly available datasets from GEO database, and the accession numbers are GSE9348, GSE32323, GSE18105 and GSE21510. All these four datasets were collected to compare gene expression in human colorectal cancer and control samples.

In order to illustrate the rationality of the idea behind our method, we compare *KL_normal* and *KL_disease* of differential pathways we detected. Our approach identified 25 significantly differential pathways on GSE32323. The specific comparison between *KL_normal* and *KL_disease* is shown in Figures 1, the horizontal axis shows the serial number of detected differential pathways, and the vertical axis means the value of KL divergence. For most detected significantl differential pathways, such as "cAMP signaling pathway","Wnt signaling pathway" and "cGMP-PKG signaling pathway", *KL_normal* is smaller than *KL_disease*, and the detailed KL-divergences of these three pathways are present in Table 2. Few disease significant pathways, such as "Pathways_in_cancer" and "Prostate_cancer", *KL_normal* is bigger than *KL_disease*. We can conclude that as to normal pathways, distributions simulated by normal gene expression fit to real distributions better. In the same way, for some disease pathways, distributions simulated by disease gene expression fit to real distributions well. Thus, this result is in accordance with our above biological hypothesis, which pathway regulatory capacity will be disturbed when organism gets to be abnormal state.

Comparing our method with GSEA, GSCA and DRAGEN, we performed these methods with the same parameters. Via searching the biological supporting documents by Google Scholar, we can get the precision of every detected result. For all presented methods, pathways with *p.value* below 0.05 are considered significantly differential, namely to compute the ratio of true significantly differential (with literature supporting) pathways in all the detected difference pathways. The detailed comparison results are shown in Table 3. For three of these expression data (GSE32323, GSE18105 and GSE9348), the results here indicate that our method performs higher precision than other three approaches.

GSEA identifies significantly differential pathways by differentially expressed (DE) genes, and it is generally accepted. In terms of precision on GSE21510, GSEA gets better detection results than our method. For assessing the extent of differentially co-expressed (DC) for a given pathway, GSCA compares correlations about all possible gene pairs for two different conditions. Thus, overlooking the true interaction of gene pairs, meanwhile expanding the difference between two phenotypes, and will identify many non-significantly differential pathways. Therefore, the precision of results by GSCA is lower. As to DRAGEN, it maps gene ID to human gene regulatory network to get interaction between genes. However, human regulatory network is extremely huge and complex. Because the network is incomplete, many genes ID can't be found the corresponding ones, then some genes with notable difference information will be lost. Consequently, DRAGEN detects few significantly differential pathways.

More seriously, DRAGEN can't identify any pathway, such as the result of performing on GSE9348.

In order to illuminate the effectiveness of our method in detecting differentially regulated pathways, by varying the cut-off value of *p.value*, we obtained a series of sensitivities and specificities on the four datasets respectively, and we were able to plot receiver operating characteristic (ROC) curves and calculate AUC scores as the area under this curve. Figures 2-5 illustrate the ROC curves of four different methods from four expression datasets (GSE32323, GSE18105, GSE21510, GSE9348), and Table 4 lists the area under curve values of the corresponding ROC curves in Figures 2-5. The horizontal axis of ROC curves shows the value of specificity, and the vertical axis means the value of sensitivity. The results of ROC curves on all four gene expression datasets show that the curves of our method climb more closely towards the top-left corner, suggesting the higher integrated performance in detecting significantly difference pathways of our method. From Table 4, we can see that AUCs of all these methods are not very high. Among 166 candidate pathways, there are almost 70 pathways have been identified relevant to colorectal cancer through literature mining. However, our method, GSEA and DRAGEN only detected a small part of differential pathways, and that may lead to low sensitivity. On the contrary, GSCA considered many pathways as statistically significant, almost above 80% of the candidate pathways, and this may lead to low specificity. Thus, the AUC, which is a trade-off between sensitivity and specificity, computed by all these four methods are not good enough. But it is worth mentioning, the AUC scores on all four gene expression datasets of our method are the highest, GSEA comes second, and the scores of GSCA and DRAGEN are not very well (most AUC scores are below 0.5), these results further support the effectiveness of our method.

Through the above experimental comparisons, our method demonstrates superior performance than the other three methods. This is partly due to the fact that detecting the probability distributions about regulatory capacity of gene pairs can provide more comprehensive view of underlying process.

Table 2. KL-divergences of pathways in different states

Pathway Name	Edge Size	KLD of Real vs Normal	KLD of Real vs Disease
cAMP signaling pathway	460	0.5773	0.9048
cGMP-PKG signaling pathway	278	0.6323	0.9348
Wnt signaling pathway	293	0.7085	1.3475

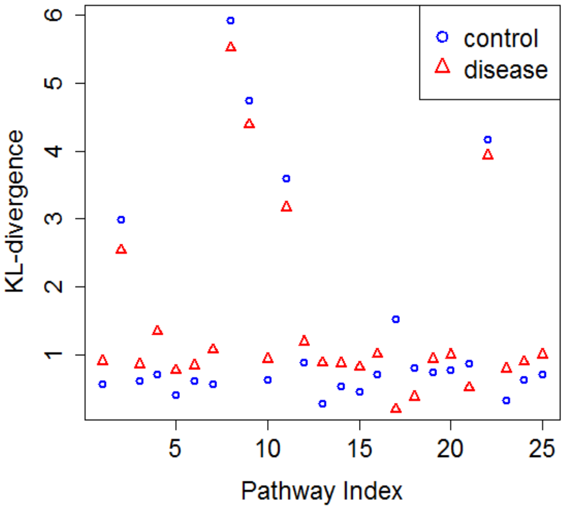

Figure 1. KL-divergences of pathways in different states

Table 3. The Precision of different methods

Data	GSEA	GSCA	DRAG-EN	Our Method
GSE32323	52.63%	47.77%	38.24%	60.00%
GSE18105	47.06%	47.77%	40.00%	61.54%
GSE21510	53.85%	47.80%	51.35%	51.72%
GSE9348	36.36%	47.62%	00.00%	73.33%

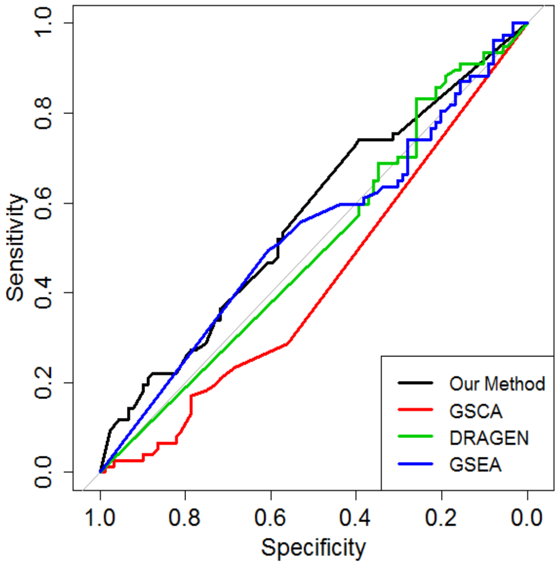

Figure 2. The comparison of ROC on GSE32323

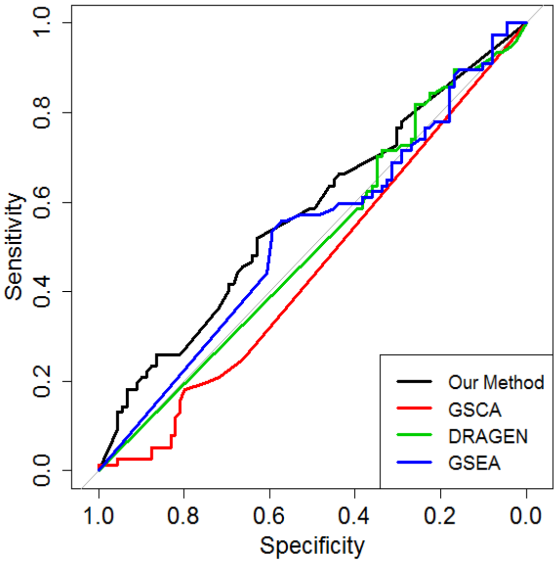

Figure 3. The comparison of ROC on GSE18105

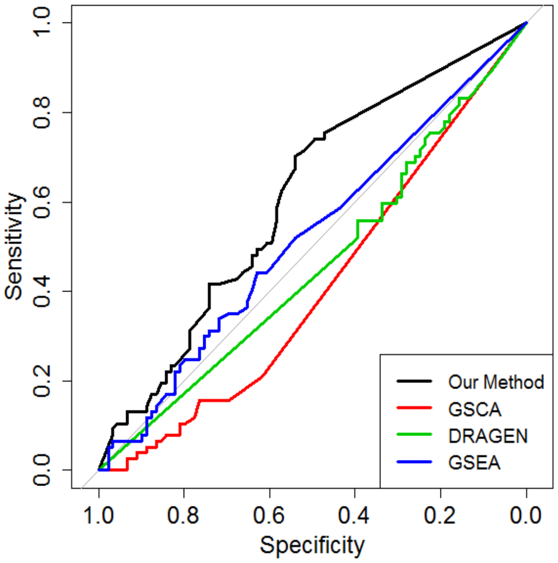

Figure 4. The comparison of ROC on GSE21510

Table 4. The area under curve values of different methods

Data	GSEA	GSCA	DRAG-	Our
			EN	Method
GSE32323	0.5302	0.4229	0.4977	0.5668
GSE18105	0.5254	0.4504	0.5012	0.5722
GSE21510	0.5209	0.4100	0.4573	0.6100
GSE9348	0.5061	0.4818	0.5000	0.6194

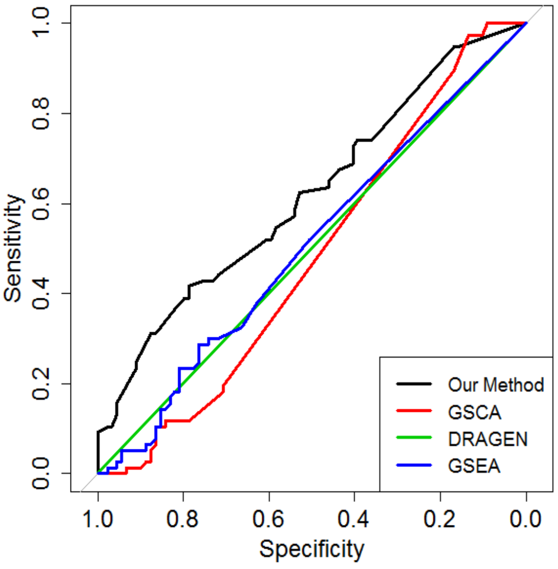

Figure 5. The comparison of ROC on GSE9348

4 Conclusion

In this paper, a new approach was proposed to identify significantly differential pathway especially between normal versus disease phenotypes. Different with other methods, we suppose the probability distribution of regulatory capacity will be changed when pathway is disturbed. Our method detects pathways based on differently regulatory (DR) genes, and it measures different regulations of pathway in a systematical way. Instead of calculating relationship of gene-gene using gene expression data alone, it integrates expression data and pathway topology information. It identifies final significantly differential pathway by comparing the distance between two fitted degrees, which is computed from different phenotype gene expression data and pathway topology structure. In the end, we implement experiments on four real microarray datasets to prove our method works. In comparison with other three previous methods, our method with higher precision is competitive to predict the true significantly disturbed pathway between normal and disease phenotypes. In addition, it is beneficial for researches on biology or medical science. In summary, the study here analysis difference of pathway from an overall perspective will provide a complementary analysis framework of pathway analysis.

Acknowledgments

We gratefully acknowledge funding support from the National Key Technologies R&D Program (2011CB910200) and Anhui Provincial Natural Science Foundation (1508085MF128).

References

[1] Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., ... & Drăghici, S. Methods and approaches in the topology-based analysis of biological pathways. Frontiers in physiology, 2013, .

[2] Khatri P, Sirota M, Butte A J. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology, 2012.

[3] Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., & Jacq, B. GOToolBox: functional analysis of gene datasets based on Gene Ontology.Genome biology, 2004..

[4] Zheng Q, Wang X J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic acids research, 2008, 36(suppl 2): W358-W363.

[5] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., & Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.Proceedings of the National Academy of Sciences of the United States of America, 2005,102(43), 15545-15550.

[6] Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F.,& Hahn, W. C. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature, 2009, 462(7269), 108-112.

[7] Fang Z, Tian W, Ji H. A network-based gene-weighting approach for pathway analysis. Cell research, 2012, 22(3): 565-580.

[8] Choi Y J, Kendziorski C. Statistical methods for gene set co-expression analysis. Bioinformatics, 2009, 25(21): 2780-2786.

[9] Rahmatallah Y, Emmert-Streib F, Glazko G. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. Bioinformatics, 2014, 30(3): 360-368.

[10] Jung S, Kim S. EDDY: a novel statistical gene set test method to detect differential genetic dependencies. Nucleic acids research, 2014.

[11] Ma S, Jiang T, Jiang R. Differential regulation enrichment analysis via the integration of transcriptional regulatory network and gene expression data. Bioinformatics, 2014.

[12] Geistlinger, L., Csaba, G., Küffner, R., Mulder, N., & Zimmer, R. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. Bioinformatics, 2011, 27(13), i366-i373.

[13] Califano A. Rewiring makes the difference. Molecular systems biology, 2011, 7(1).

[14] Kullback S, Leibler R A. On information and sufficiency. The annals of mathematical statistics, 1951, 79-86.

[15] Edgar R, Domrachev M, Lash A E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 2002, 30(1): 207-210.

[16] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 2000, 28(1): 27-30.

[17] Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., & Valencia, A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics, 2012, 28(18), i451-i457..

[18] Liu, W., Li, C., Xu, Y., Yang, H., Yao, Q., Han, J.,. & Li, X.. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. Bioinformatics, 2013, 29(17), 2169-2177.